An analysis of the spectrum of the discontinuous Galerkin method

Lilia Krivodonova[a] and Ruibin Qin[b]
Department of Applied Mathematics
University of Waterloo
200 University Ave West
Waterloo, ON, N2L 3G1
Canada
a) corresponding author, lgk@math.uwaterloo.ca, 1-519-888-4567 ext 38138
b) rqin@math.uwaterloo.ca

Abstract

We derive explicit expressions for the eigenvalues (spectrum) of the discontinuous Galerkin spatial discretization applied to the linear advection equation. We show that the eigenvalues are related to the subdiagonal $[p/p+1]$ Padé approximation of exp(-z) when p-th degree basis functions are used. We derive an upper bound on the eigenvalue with the largest magnitude as $(p+1)(p+2)$. We demonstrate that this bound is not tight and prove that the asymptotic growth rate of the spectral radius is slower than quadratic in p. We also analyze the behavior of the spectrum near the imaginary axis to demonstrate that the spectral curves approach the imaginary axis although there are no purely imaginary eigenvalues.

1 Introduction

In this paper we derive explicit expressions for the eigenvalues (spectrum) of the semi-discrete discontinuous Galerkin (DG) method applied to the one-dimensional linear advection equation. The DG spatial discretization results in a linear system of ODEs

$$\frac{d}{dt}\mathbf{c} = \frac{a}{\Delta x} L\mathbf{c} \tag{1}$$

for $(p+1)N$ degrees of freedom $\mathbf{c}$ on an N element uniform mesh with p-th degree approximation in space. Here, a is the wave speed and Δx is the cell size. We show that for a discretization with the upwind flux and periodic boundary conditions, the eigenvalues of L are given by $f_{p+1}(\lambda) = exp(\frac{2\pi i}{N} j)$, $j = 0, 1, \ldots, N-1$, where $f_{p+1}(z)$ is the subdiagonal $[p/p+1]$ Padé approximant of $exp(-z)$. We also demonstrate that the eigenvectors of L are related to N-th roots of unity.

A direct application of the eigenvalue analysis is to the linear stability of the fully discrete scheme. Equation (1) is usually integrated in time using a suitable ODE solver. Thus, the necessary condition for the stability of the method is to require the time step Δt to be small enough so that the full spectrum of $\frac{a\Delta t}{\Delta x}L$ fits inside the absolute stability region of the chosen time integration scheme. The eigenvalues of L can be computed using a linear algebra software which has been done for a variety of combinations of spatial orders and time integration schemes [7, 13]. However, the analytical form of the eigenvalues has not been previously known. It is interesting from a purely theoretical point of view and can also be used to get further insight into the DG method. We use it to improve the CFL number by manipulating the scheme so that the spectrum of L is shrunk [4]. This is achieved by constructing a different rational approximant of $exp(-z)$ which seeks to preserve the order of accuracy in the L^2 norm.

A linear stability analysis of (1) arising from a low order DG spatial discretization and Runge-Kutta time integration was previously performed in [6, 5] and, more recently, in [16] for two-dimensional problems. It was shown that the DGM with $p > 0$ is not stable with a fixed CFL number when the forward Euler time integration is used [5]. This is caused by the eigenvalues of (1) being located very close to the imaginary axis which is not included in the stability region of the forward Euler method. It was proven in [6] that the DG method with $p = 1$ and the second order Runge-Kutta scheme is L^2 stable with the CFL number equal to $1/3$. It was further hypothesized there that a coupling of a pth degree DG scheme with a $(p+1)$st order RK scheme is stable under a CFL condition $1/(2p+1)$. In recent years, the DGM has been used with a variety of explicit time integration schemes, such as Adams-Bashforth [8], strong-stability preserving schemes [9], low storage RK schemes [13]. In this view, the universal CFL number seems to be of less importance.

Using the obtained expressions for the eigenvalues, we analyze the asymptotic behavior of the spectrum as the order of approximation p goes to infinity. The real eigenvalue, which is conjectured to be the largest in magnitude, and the real component of complex eigenvalues is shown to be bounded from below by $-(p+1)(p+2)$ for any p. However, we prove that the actual growth rate of the size of the largest eigenvalue is slower than quadratic. Numerical experiments indicate that $-1.5(p+1)^{1.75}$ is an upper bound on the eigenvalues. The least square fitting gives a growth rate of about $1.4(p+1)^{1.78}$ for $p < 100$. We also demonstrate that although the curves $|f_{p+1}(z)| = 1$ move closer to the imaginary axis as p increases there are no purely imaginary eigenvalues for any p.

A connection between the DG method and the Padé approximants has been observed previously. In [17], Le Saint and Raviart showed that the absolute stability region of the discontinuous Galerkin method used to solve an ODE is given by $|R(\lambda h)| \leq 1$, where $R(z)$ is the $[p/p+1]$ Padé approximant of $exp(z)$. In [14], Hu and Atkins studied the dispersion properties of the DG scheme applied to the scalar advection equation in one dimension. They showed that for the physical mode, the numerical dispersion relation is accurate to $(k\Delta x)^{2p+2}$, where k is the wavenumber and $k\Delta x$ is small. Their reasoning was founded on the conjecture that certain polynomials involved in the analysis are related to $[p+1/p]$ Padé approximation of $exp(z)$. An extended analysis of the dispersion and dissipation errors were given by Ainsworth in [2]. It was demonstrated there that the numerical wave speed $\tilde{k}$ satisfies the relation $f_{p+1}(-i\Delta xk) = exp(i\Delta x\tilde{k})$. The proof is based on a demonstration that DG solutions satisfy a certain eigenvalue problem conjectured in [14]. In Theorem 1 we show how this eigenvalue problem arises from the characteristic polynomial of L.

The $[p/p+1]$ and $[p+1/p]$ Padé approximants of $exp(z)$ are $O(z^{2p+2})$ accurate for small z. This explains the excellent dispersion and dissipation properties of the DGM which were called "superconvergent" in [14, 2]. This makes the scheme very suitable for wave propagation problems especially ones requiring long time integration. However, from our analysis it follows that the same approximants are involved in defining the spectrum of the semi-discrete method and, in this sense, are responsible for a severe time step restriction associated with the DGM. The small CFL number is frequently quoted as an disadvantage of the DGM. It makes the method, especially for low p and nonlinear problems, more expensive when compared to schemes that are able to maintain the CFL close to unity, e.g. finite volume schemes.

The rest of this paper is organized as follows. We begin by deriving the discontinuous Galerkin formulation of the model problem with the aim to obtain a general form of the resulting systems of ODEs. In Section 3, we derive the equations that describe the eigenvalues and eigenvectors of the spatial discretization and prove our main result, i.e the relation between the characteristic polynomial of L and Padé approximants. Section 4 contains an analysis of the distribution of eigenvalues and the growth speed of the eigenvalue of the largest modulus. Finally, conclusions and discussions are provided in Section 5.

2 Discontinuous Galerkin discretization

We consider the one-dimensional linear advection equation

$$u_t + au_x = 0 \tag{2}$$

subject to appropriate initial and periodic boundary conditions on interval I, $a > 0$. The domain is discretized uniformly into mesh elements $I_j = [x_{j-1}, x_j]$ of size Δx, $j = 1, 2, ..., N$. The discontinuous Galerkin spatial discretization on cell I_j with the upwind numerical flux is obtained by approximating u by $U_j \in \mathcal{P}_p$, multiplying (2) by a test function $V \in \mathcal{P}_p$, integrating the result on I_j while using integration by parts once

$$\frac{d}{dt}\int_{x_{j-1}}^{x_j} U_j V\,dx + aU_j(x_j)V(x_j) - aU_{j-1}(x_{j-1})V(x_{j-1}) - a\int_{x_{j-1}}^{x_j} U_j V'\,dx = 0, \quad \forall V \in \mathcal{P}_p. \tag{3}$$

$\mathcal{P}_p$ is a finite dimensional space of polynomials of degree up to p. Transforming $[x_{j-1}, x_j]$ to the canonical element $[-1, 1]$ by a linear mapping

$$x(\xi) = \frac{x_{j-1} + x_j}{2} + \frac{\Delta x}{2}\xi \tag{4}$$

yields

$$\frac{\Delta x}{2}\frac{d}{dt}\int_{-1}^{1} U_j V\,d\xi + aU_j(1)V(1) - aU_{j-1}(1)V(-1) - a\int_{-1}^{1} U_j V'\,d\xi = 0, \quad \forall V \in \mathcal{P}_p. \tag{5}$$

We choose the Legendre polynomials as the basis for the finite element space $\mathcal{P}_p$. Recall [1], that the Legendre polynomials $P_k(\xi)$, $k = 0, 1, 2, \ldots$, form an orthogonal system on $[-1, 1]$

$$\int_{-1}^{1} P_k P_i\,d\xi = \frac{2}{2k+1}\delta_{ki}, \tag{6}$$

where δ_{ki} is the Kroneker delta. With the chosen normalization (6), the values of the basis functions at the end points of the interval $[-1,1]$ are [1]

$$P_k(1) = 1, \qquad P_k(-1) = (-1)^k. \tag{7}$$

The numerical solution can be written in terms of the basis as

$$U_j = \sum_{i=0}^{p} c_{ji} P_i, \tag{8}$$

where c_{ji} is a function of time t. Substituting (8) into (5), choosing $V = P_k$, $k = 0, 1, \ldots, p$, and using (7) and (6) results in

$$\frac{\Delta x}{2k+1}\dot{c}_{jk} = -a\left(\sum_{i=0}^{p} c_{ji} - (-1)^k \sum_{i=0}^{p} c_{j-1,i}\right) + a\int_{-1}^{1}\left(\sum_{i=0}^{p} c_{ji}P_i\right)P_k'\, d\xi, \quad k = 0, 1, \ldots, p, \tag{9}$$

where the dot in $\dot{c}_{jk}$ represents differentiation with respect to t. Collecting common terms of c_{ji} results in

$$\dot{c}_{jk} = a\frac{2k+1}{\Delta x}\left[(-1)^k \sum_{i=0}^{p} c_{j-1,i} + \sum_{i=0}^{p}\left(\int_{-1}^{1} P_i P_k'\, d\xi - 1\right)c_{ji}\right], \quad k = 0, 1, \ldots, p. \tag{10}$$

This can be written in a vector form as

$$\dot{c}_{jk} = a\frac{2k+1}{\Delta x}\left((-1)^k[1,1,...,1]\mathbf{c}_{j-1} + [\int_{-1}^{1} P_0 P_k' d\xi - 1, ..., \int_{-1}^{1} P_p P_k' d\xi - 1]\mathbf{c}_j\right), \tag{11}$$

where $\mathbf{c}_j = [c_{j0}, c_{j1}, \ldots, c_{jp}]^T$ and $\mathbf{c}_{j-1}$ is defined similarly. Combining cell solution-coefficient vectors into a global vector $\mathbf{c} = [\mathbf{c}_0^T, \mathbf{c}_1^T, \ldots, \mathbf{c}_p^T]^T$, equation (11) can be written as

$$\dot{\mathbf{c}} = \frac{a}{\Delta x} L\mathbf{c}. \tag{12}$$

With periodic boundary conditions, L is a block matrix of the form

$$L = \begin{bmatrix} A_n & 0 & 0 & \ldots & 0 & 0 & D_n \\ D_n & A_n & 0 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & D_n & A_n \end{bmatrix}, \tag{13}$$

where D_n and A_n are $n \times n$ matrices, $n = p+1$. For approximation of order p, there are $p+1$ basis functions, so the size of each block is $(p+1)\times(p+1)$. In the following discussion this notation of n is consistent and n can always be replaced by $p+1$. In the matrix L,

$$D_n = \begin{bmatrix} 1 & \ldots & 1 \\ -3 & \ldots & -3 \\ \vdots & & \vdots \\ (-1)^{n-1}(2n-1) & \ldots & (-1)^{n-1}(2n-1) \end{bmatrix}, \tag{14}$$

$$A_n = \begin{bmatrix} \int_{-1}^{1} P_0 P_0' d\xi - 1 & \dots & \int_{-1}^{1} P_{n-1} P_0' d\xi - 1 \\ 3\left(\int_{-1}^{1} P_0 P_1' d\xi - 1\right) & \dots & 3\left(\int_{-1}^{1} P_{n-1} P_1' d\xi - 1\right) \\ \vdots & & \vdots \\ (2n-1)\left(\int_{-1}^{1} P_0 P_{n-1}' d\xi - 1\right) & \dots & (2n-1)\left(\int_{-1}^{1} P_{n-1} P_{n-1}' d\xi - 1\right) \end{bmatrix}, \quad (15)$$

or

$$A_n = (a_{ij}) = \left((2i-1)(\int_{-1}^{1} P_{j-1} P_{i-1}' d\xi - 1) \right). \quad (16)$$

Noticing that the derivatives of the Legendre polynomials satisfy [1]

$$(2k+1)P_k = P_{k+1}' - P_{k-1}', \quad (17)$$

we derive

$$P_{k+1}' = (2k+1)P_k + (2(k-2)+1)P_{k-2} + (2(k-4)+1)P_{k-4} + \dots. \quad (18)$$

We use (18) with the orthogonality property of the Legendre polynomials (6) to simplify the integrals in A_n. We obtain

$$\int_{-1}^{1} P_i P_k' d\xi = \begin{cases} 0, & k \leqslant i, \\ 2, & k > i, \text{ and } (k-i) \equiv 1 \pmod 2, \\ 0, & k > i, \text{ and } (k-i) \equiv 0 \pmod 2. \end{cases} \quad (19)$$

Thus, A_n can be simplified as

$$A_n = -\begin{bmatrix} a_1 & a_1 & a_1 & \cdots & a_1 & a_1 \\ -a_2 & a_2 & a_2 & \cdots & a_2 & a_2 \\ a_3 & -a_3 & a_3 & \cdots & a_3 & a_3 \\ -a_4 & a_4 & -a_4 & \cdots & a_4 & a_4 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ (-1)^{n-2}a_{n-1} & (-1)^{n-3}a_{n-1} & (-1)^{n-4}a_{n-1} & \cdots & a_{n-1} & a_{n-1} \\ (-1)^{n-1}a_n & (-1)^{n-2}a_n & (-1)^{n-3}a_n & \cdots & -a_n & a_n \end{bmatrix}, \quad (20)$$

where $a_i = 2i-1,\ i = 1,2,\dots,n$.

3 Characteristic polynomial of L and the Padé approximant

Next, we derive an expression for the eigenvalues of L. λ is an eigenvalue of L if it satisfies

$$\begin{bmatrix} A_n & 0 & 0 & \dots & 0 & 0 & D_n \\ D_n & A_n & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & D_n & A_n \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_N \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_N \end{bmatrix}, \quad (21)$$

where $\mathbf{v}^T = [\mathbf{v}_1^T, \mathbf{v}_2^T, \dots, \mathbf{v}_N^T]$ is the corresponding eigenvector and its components $\mathbf{v}_j,\ j = 1,2,\dots,N$, are column vectors of length $n = p+1$. Equivalently, we can write equation (21) as

$$D_n \mathbf{v}_{j-1} + A_n \mathbf{v}_j = \lambda \mathbf{v}_j, \quad j = 1,2,\dots,N, \quad (22)$$

with an understanding that $\mathbf{v}_0 = \mathbf{v}_N$. We express D_n defined by (14) as an outer product $D_n = \mathbf{r}_n[1,1,...,1]$, where $\mathbf{r}_n = [1,-3,\dots,(-1)^{n-1}(2n-1)]^T$. Then (22) can be rewritten as

$$\mathbf{r}_n[1,1,...,1]\mathbf{v}_{j-1} = (\lambda I - A_n)\mathbf{v}_j. \tag{23}$$

Introducing a new variable $S_j = [1,1,...,1]\cdot\mathbf{v}_j$, we write (23) as

$$S_{j-1}\mathbf{r}_n = (\lambda I - A_n)\mathbf{v}_j. \tag{24}$$

Multiplying both sides of (24) by $[1,1,...,1](\lambda I - A_n)^{-1}$ yields

$$S_j = S_{j-1}[1,1,...,1](\lambda I - A_n)^{-1}\mathbf{r}_n. \tag{25}$$

Let

$$f_n(\lambda) = [1,1,...,1](\lambda I - A_n)^{-1}\mathbf{r}_n. \tag{26}$$

Then, (25) results in a recursive formula

$$S_j = f_n(\lambda)S_{j-1}. \tag{27}$$

Expansion of (27) starting with $j = N$ gives

$$S_N = f_n^{N-1}(\lambda)S_1. \tag{28}$$

Finally, taking into account periodicity of the boundary conditions, we obtain $S_N = f_n^N(\lambda)S_N$. This implies

$$f_n^N(\lambda) = 1. \tag{29}$$

Then, the eigenvalues of L are the roots of the equations

$$f_n(\lambda) = \omega_j, \quad \omega_j = e^{\frac{2\pi i}{N}j}, \quad j = 0,1,2,\dots,N-1. \tag{30}$$

Eigenvectors. For completeness of this discussion, we derive the eigenvectors of matrix L. Since L is a block circulant matrix, we look for eigenvectors $\mathbf{v}$ in the form $[\tilde{\mathbf{v}}^T, \omega_k\tilde{\mathbf{v}}^T, \dots, \omega_k^{N-1}\tilde{\mathbf{v}}^T]^T$. Substituting $\mathbf{v}$ into (21) gives

$$\omega_k^{j-1}D_n\tilde{\mathbf{v}} + \omega_k^j A_n\tilde{\mathbf{v}} = \lambda_k\omega_k^j\tilde{\mathbf{v}}, \quad 1 \leqslant j \leqslant N, \tag{31}$$

or

$$(\omega_k\lambda_k I - \omega_k A_n - D_n)\tilde{\mathbf{v}} = 0, \tag{32}$$

where λ_k is one of the roots of $f_n(\lambda) = \omega_k$. $\tilde{\mathbf{v}}$ can be easily obtained by solving the linear system (32). The solutions are not particularly illuminating and we do not report them. Figure 1 shows the periodic property of the components of the eigenvectors. We plot one of the two eigenvectors corresponding to $k = 4$ (left) and $k = 17$ (right). In figure 1 each point corresponds to an entry in $\mathbf{v}$. The entries of the eigenvectors represent sampling of a scaled unit circle at N or, if N/k is an integer, N/k points. One of the circles in Figure 1, left and right, corresponds to the first entry of $\omega_k^j\tilde{\mathbf{v}}$, $j = 1,2,\dots,N-1$, and the other to the second entry of $\omega_k^j\tilde{\mathbf{v}}$. The line connecting two points represents two consecutive entries of $\tilde{\mathbf{v}}$ and, thus, the shift in sampling between the two components of each $\omega_k^j\tilde{\mathbf{v}}$, $k = 1,2,\dots,N-1$,.

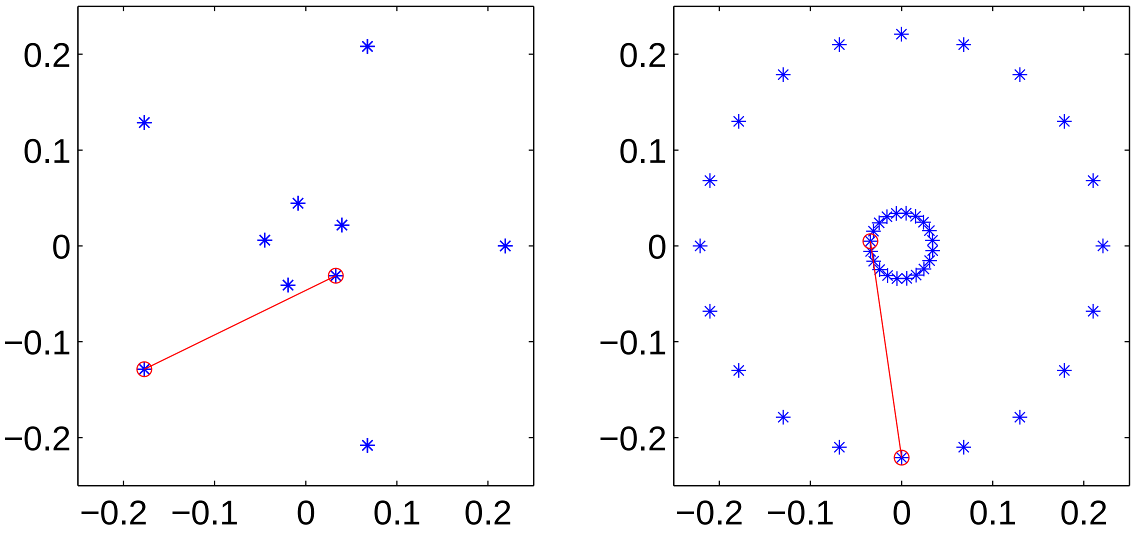

Figure 1: Eigenvectors of L with $N = 20$, $p = 1$, and $k = 4$ (left) and $k = 17$ (right). Each point in plots correspond to an entry in an eigenvector. The two points connected by a line show the first two entries of $\tilde{\mathbf{v}}$.

Padé approximants. In the theorem that follows we will demonstrate that function $f_n(z)$ is the $[n-1/n]$ Padé approximant of e^{-z}. Recall [1], that the Padé approximant is a rational approximation to a given function. Let us suppose that we are given the Taylor expansion of a function $g(z) = \sum_{i=0}^{\infty} c_i z^i$. A Padé approximant is a fraction

$$[L/M] = \frac{a_0 + a_1 z + \cdots + a_L z^L}{b_0 + b_1 z + \cdots + b_M z^M}, \tag{33}$$

that satisfies

$$\sum_{i=0}^{\infty} c_i z^i = \frac{a_0 + a_1 z + \cdots + a_L z^L}{b_0 + b_1 z + \cdots + b_M z^M} + O(z^{L+M+1}). \tag{34}$$

Coefficients $a_0, a_1, \ldots, a_L$, and $b_0, b_1, \ldots, b_M$ are uniquely defined by $c_0, c_1, \ldots$, if a_0 is fixed. It is a common practice to display the approximants in a table, which is called the Padé table. A part of the Padé table of e^z is illustrated in Appendix (Table 2) as an example.

The Padé approximants of the exponential function e^z are shown to be given by the following formula for non-negative integers p, q [3]

$$[p/q]_{exp(z)} = \frac{{}_1F_1(-p, -p-q, z)}{{}_1F_1(-q, -p-q, -z)}, \tag{35}$$

where ${}_1F_1$ denotes the confluent hypergeometric function defined by the series [1]

$$ {}_1F_1(a, b, z) = 1 + \frac{a}{b} z + \frac{a}{b} \frac{a+1}{b+1} \frac{z^2}{2!} + \frac{a}{b} \frac{a+1}{b+1} \frac{a+2}{b+2} \frac{z^3}{3!} + \cdots. \tag{36}$$

When a, b are negative integers and $b \leqslant a$, ${}_1F_1(a,b,z)$ is a finite sum which is a polynomial of degree $|a|$. Using the Pochhammer's symbol, $(a)_k = a(a+1)\cdots(a+k-1)$ and $(a)_0 = 1$, we can rewrite (36) in a compact form

$$ {}_1F_1(a,b,z) = \sum_{k=0}^{\infty} \frac{(a)_k}{(b)_k} \frac{z^k}{k!}. \tag{37}$$

In the following theorem, we state our main result.

Theorem 1. *If A_n is an $n \times n$ matrix given by (20), and $f_n(z) = (1,...,1)(zI - A_n)^{-1}\mathbf{r}_n$, where $\mathbf{r}_n = \left[1, -3, \cdots, (-1)^{n-1}(2n-1)\right]^T$, then*

$$ f_n(z) = \frac{{}_1F_1(-n+1, -2n+1, -z)}{{}_1F_1(-n, -2n+1, z)}, \tag{38}$$

which is the $[n-1/n]$ Padé approximant of e^{-z}.

In order to prove the theorem, we will need to establish three auxiliary results which are proved in the following three lemmas. We start by introducing additional notation.

Definition 1.

- *$\tilde{A}_n$: a matrix defined as $\tilde{A}_n = zI - A_n$.*
- *$M_{n,i}$: the (n,i) minor of $\tilde{A}_n$, i.e. the determinant of the $(n-1) \times (n-1)$ matrix obtained by elimination of the n-th row and i-th column of $\tilde{A}_n$, $i = 1, 2, \ldots, n$.*
- *$\tilde{A}_n^j$: the $n \times n$ matrix obtained by replacing the j-th column of $\tilde{A}_n$ with $\mathbf{r}_n$, $j = 1, 2, ..., n$.*
- *M_n^j: the determinant of $\tilde{A}_n^j$, $j = 1, 2, \ldots, n$.*
- *$M_{n,i}^j$: the (n,i) minor of the matrix $\tilde{A}_n^j$, $i, j = 1, 2, \ldots, n$.*

We also introduce two sequences of polynomials which are essential to our proofs

$$ \begin{cases} Q_n(z) &= (a_n + z)Q_{n-1}(z) + a_n R_{n-1}(z), \\ R_n(z) &= a_n Q_{n-1}(z) + (a_n - z)R_{n-1}(z), \end{cases} \tag{39}$$

where $Q_1(z) = a_1 + z$, $R_1(z) = a_1$, and $a_n = 2n - 1$. As an example, $Q_n(z)$ and $R_n(z)$ for small n are listed in Table 3 in Appendix. Note that while Q_n is a polynomial of degree n, R_n is a polynomial of degree $n-1$. We will show that Q_n and R_n are proportional to the hypergeometric functions appearing in (38) and give an alternative expression for $f_n(z)$

$$ f_n(z) = \frac{R_n(-z)}{Q_n(z)}. \tag{40}$$

Thus, (39) is a recursive formula for generating $[p/p+1]$ and $[p+1/p]$ (sub- and superdiagonal) Padé approximants for $exp(-z)$.

We start with Lemma 1 which relates $Q_n(z)$ and $R_n(z)$ to the determinant of $\tilde{A}_n$ and its minors.

Lemma 1. *Let $Q_n(z)$ and $R_n(z)$ be defined by (39). Then*

$$Q_n(z) = \det(\tilde{A}_n), \tag{41a}$$

$$R_n(z) = \sum_{i=1}^{n} M_{n+1,i}. \tag{41b}$$

Proof. We will use an induction argument to prove (41). By Definition 1 and (39), $Q_1(z) = a_1 + z = \det(\tilde{A}_1)$, and $R_1(z) = a_1 = M_{2,1}$. This establishes the base of the induction. We assume that (41) holds for $Q_n(z), R_n(z)$, and we will prove that it is valid for $Q_{n+1}(z), R_{n+1}(z)$.

Applying the cofactor expansion along the $(n+1)$-th row of $\det(\tilde{A}_{n+1})$ while noticing that $M_{n+1,n+1} = \det(\tilde{A}_n)$ yields

$$\begin{aligned} \det(\tilde{A}_{n+1}) &= \sum_{i=1}^{n} (-1)^{n+1-i} a_{n+1} (-1)^{n+1+i} M_{n+1,i} + (a_{n+1}+z) M_{n+1,n+1} \\ &= a_{n+1} \sum_{i=1}^{n} M_{n+1,i} + (a_{n+1}+z) \det(\tilde{A}_n) \\ &= (a_{n+1}+z) Q_n(z) + a_{n+1} R_n(z) = Q_{n+1}(z). \end{aligned} \tag{42}$$

This proves the recursion (41a) for $Q_n(z)$.

Next, we prove (41b) for $R_{n+1}(z) = \sum_{i=1}^{n+1} M_{n+2,i}$. We begin by relating $M_{n+2,i}$ to $M_{n+1,i}$, $i < n+1$. In (43), we write an explicit expression for $M_{n+2,i}$, then subtract the $n+1$st column from the nth column and compute the determinant by a cofactor expansion based on the nth column,

$$M_{n+2,i} = \begin{vmatrix} a_1+z & \cdots & a_1 & a_1 \\ -a_2 & \cdots & a_2 & a_2 \\ \vdots & & \vdots & \vdots \\ (-1)^n a_{n+1} & \cdots & a_{n+1}+z & a_{n+1} \end{vmatrix} = \begin{vmatrix} a_1+z & \cdots & 0 & a_1 \\ -a_2 & \cdots & 0 & a_2 \\ \vdots & & \vdots & \vdots \\ (-1)^n a_{n+1} & \cdots & z & a_{n+1} \end{vmatrix} = -z M_{n+1,i}. \tag{43}$$

Similarly, for $i = n+1$, a cofactor expansion based on the last row yields

$$\begin{aligned} M_{n+2,n+1} &= \sum_{i=1}^{n} \left[(-1)^{n+1-i} (-1)^{n+1+i} a_{n+1} M_{n+1,i} \right] + a_{n+1} M_{n+1,n+1} \\ &= a_{n+1} \sum_{i=1}^{n} M_{n+1,i} + a_{n+1} M_{n+1,n+1}. \end{aligned} \tag{44}$$

Thus,

$$\begin{aligned} \sum_{i=1}^{n+1} M_{n+2,i} &= \sum_{i=1}^{n} (-z) M_{n+1,i} + a_{n+1} \sum_{i=1}^{n} M_{n+1,i} + a_{n+1} M_{n+1,n+1} \\ &= (a_{n+1} - z) \sum_{i=1}^{n} M_{n+1,i} + a_{n+1} M_{n+1,n+1} \\ &= (a_{n+1} - z) R_n(z) + a_{n+1} Q_n(z) \\ &= R_{n+1}(z). \end{aligned} \tag{45}$$

This completes the proof. □

Lemma 2 relates $R_n(z)$ and $Q_n(z)$ to the determinant of $\tilde{A}_n^j$ and its minors.

Lemma 2. *Let $R_n(z),\ Q_n(z)$ be defined by (39). Then,*

$$R_n(-z) = \sum_{j=1}^{n} M_n^j, \tag{46a}$$

$$Q_n(-z) = \sum_{j=1}^{n+1} \left[\sum_{\substack{i=1 \\ i\neq j}}^{n} M_{n+1,i}^j + (-1)^{j-1} M_{n+1,j}^j \right]. \tag{46b}$$

Proof. The case $n=1$ is satisfied trivially by the involved variables given by Definition 1 and (39). We assume that (46) is true for n, and we will prove it is also true for $n+1$.

Applying cofactor expansion to $M_{n+1}^j,\ j=1,\ldots,n$ along the last row gives

$$\begin{aligned} M_{n+1}^j &= \sum_{\substack{i=1 \\ i\neq j}}^{n} \left[(-1)^{n+1-i} a_{n+1} (-1)^{n+1+i} M_{n+1,i}^j \right] + (-1)^n a_{n+1} (-1)^{n+1+j} M_{n+1,j}^j + (a_{n+1}+z) M_{n+1,n+1}^j \\ &= a_{n+1} \sum_{\substack{i=1 \\ i\neq j}}^{n} M_{n+1,i}^j + (-1)^{j-1} a_{n+1} M_{n+1,j}^j + (a_{n+1}+z) M_{n+1,n+1}^j. \end{aligned} \tag{47}$$

Similarly, applying cofactor expansion to M_{n+1}^{n+1} along the last row gives

$$\begin{aligned} M_{n+1}^{n+1} &= \sum_{i=1}^{n} (-1)^{n+1-i} a_{n+1} (-1)^{n+1+i} M_{n+1,i}^{n+1} + (-1)^n a_{n+1} M_{n+1,n+1}^{n+1} \\ &= a_{n+1} \sum_{i=1}^{n} M_{n+1,i}^{n+1} + (-1)^n a_{n+1} M_{n+1,n+1}^{n+1}. \end{aligned} \tag{48}$$

Since $M_{n+1,n+1}^j = M_n^j$, we can write

$$\begin{aligned} \sum_{j=1}^{n+1} M_{n+1}^j &= (a_{n+1}+z) \sum_{j=1}^{n} M_n^j + a_{n+1} \sum_{j=1}^{n+1} \left[\sum_{\substack{i=1 \\ i\neq j}}^{n} M_{n+1,i}^j + (-1)^{j-1} M_{n+1,j}^j \right] \\ &= (a_{n+1}+z) R_n(-z) + a_{n+1} Q_n(-z) \\ &= R_{n+1}(-z). \end{aligned} \tag{49}$$

This proves (46a).

We split the proof of (46b) into 2 parts: $j=1,2,\ldots,n$ and $j=n+1,n+2$. For $j=1,\ldots,n$, using an argument similar to one employed in (43), we can derive $M_{n+2,i}^j = (-z) M_{n+1,i}^j,\ i =$

$1,2,\ldots,n$. This with a cofactor expansion on the last row of $M^j_{n+2,n+1}$ gives

$$\begin{aligned}\sum_{\substack{i=1\\i\neq j}}^{n+1} M^j_{n+2,i}+(-1)^{j-1}M^j_{n+2,j} &= \sum_{\substack{i=1\\i\neq j}}^{n} M^j_{n+2,i}+(-1)^{j-1}M^j_{n+2,j}+M^j_{n+2,n+1}\\ &= (-z)\left[\sum_{\substack{i=1\\i\neq j}}^{n} M^j_{n+1,i}+(-1)^{j-1}M^j_{n+1,j}\right]\\ &\quad +a_{n+1}\left[\sum_{\substack{i=1\\i\neq j}}^{n} M^j_{n+1,i}+(-1)^{j-1}M^j_{n+1,j}\right]+a_{n+1}M^j_{n+1,n+1}.\end{aligned} \tag{50}$$

For $j=n+1,n+2$, we switch the last two columns of $M^{n+2}_{n+2,i}$ to obtain

$$\begin{aligned}M^{n+2}_{n+2,i} &= \begin{vmatrix} a_1+z & \cdots & a_1 & a_1\\ -a_2 & \cdots & a_2 & -a_2\\ \vdots & & \vdots & \vdots\\ (-1)^n a_{n+1} & \cdots & a_{n+1}+z & (-1)^n a_{n+1}\end{vmatrix}\\ &= -\begin{vmatrix} a_1+z & \cdots & a_1 & a_1\\ -a_2 & \cdots & -a_2 & a_2\\ \vdots & & \vdots & \vdots\\ (-1)^n a_{n+1} & \cdots & (-1)^n a_{n+1} & a_{n+1}+z\end{vmatrix}\end{aligned} \tag{51}$$

Comparing $-M^{n+2}_{n+2,i}$ with $M^{n+1}_{n+2,i}$ reveals that the entries in the determinants are identical except for the $(n+1,n+1)$ element, which is $(a_{n+1}+z)$ in $-M^{n+2}_{n+2,i}$ and a_{n+1} in $M^{n+1}_{n+2,i}$. Expanding the determinants along the last rows of $M^{n+1}_{n+2,i}$ and $M^{n+2}_{n+2,i}$ and adding up the results, we have

$$M^{n+1}_{n+2,i}+M^{n+2}_{n+2,i}=(-z)M^{n+1}_{n+1,i},\ i=1,2,\ldots,n. \tag{52}$$

A similar observation gives

$$M^{n+1}_{n+2,n+1}+M^{n+2}_{n+2,n+2}=(-z)M^{n+1}_{n+1,n+1}. \tag{53}$$

Combining (52) and (53) and using a cofactor expansion on $M^{n+2}_{n+2,n+1}$ along the last row, we obtain

$$\begin{aligned}&\sum_{j=n+1}^{n+2}\sum_{i=1}^{n} M^j_{n+2,i}+(-1)^n M^{n+1}_{n+2,n+1}+M^{n+2}_{n+2,n+1}+(-1)^{n+1}M^{n+2}_{n+2,n+2}\\ =\ &(-z)\sum_{i=1}^{n} M^{n+1}_{n+1,i}+(-1)^n(a_{n+1}-z)M^{n+1}_{n+1,n+1}+a_{n+1}\sum_{i=1}^{n} M^{n+1}_{n+1,i}.\end{aligned} \tag{54}$$

Finally, combining (50) and (54) yields the result

$$\begin{aligned}
\sum_{j=1}^{n+2}\left[\sum_{\substack{i=1\\i\neq j}}^{n+1} M_{n+2,i}^{j}+(-1)^{j-1}M_{n+2,j}^{j}\right] &= (a_{n+1}-z)\sum_{j=1}^{n+1}\left[\sum_{\substack{i=1\\i\neq j}}^{n} M_{n+1,i}^{j}+(-1)^{j-1}M_{n+1,j}^{j}\right]\\
&\quad +a_{n+1}\sum_{j=1}^{n} M_{n+1,n+1}^{j}\\
&= (a_{n+1}-z)Q_n(-z)+a_{n+1}R_n(-z)\\
&= Q_{n+1}(-z),
\end{aligned} \tag{55}$$

which completes the proof. □

Lemma 3 relates polynomials $Q_n(z)$ and $R_n(z)$ to the confluent hypergeometric functions.

Lemma 3. *Let $Q_n(z)$ and $R_n(z)$ be polynomials defined by (39). Then,*

$$Q_n(z)=\prod_{i=1}^{n} a_i 2^{n-1} {}_1F_1(-n,-2n+1,z), \tag{56a}$$

$$R_n(z)=\prod_{i=1}^{n} a_i 2^{n-1} {}_1F_1(-n+1,-2n+1,z). \tag{56b}$$

Proof. When $n=1$, (56) is validated by Definition 1 and (39). Assuming that (56) is true for $Q_n(z), R_n(z)$, we will show it is also true for $Q_{n+1}(z), R_{n+1}(z)$. We start with (56a). By the recurrence relation (39) and the assumption that (56) is true for $Q_n(z), R_n(z)$, we have

$$\begin{aligned}
Q_{n+1}(z) &= (a_{n+1}+z)Q_n(z)+a_{n+1}R_n(z)\\
&= (a_{n+1}+z)\prod_{i=1}^{n} a_i 2^{n-1} {}_1F_1(-n,-2n+1,z)+a_{n+1}\prod_{i=1}^{n} a_i 2^{n-1} {}_1F_1(-n+1,-2n+1,z)\\
&= \prod_{i=1}^{n} a_i 2^{n-1}\left[(a_{n+1}+z){}_1F_1(-n,-2n+1,z)+a_{n+1}{}_1F_1(-n+1,-2n+1,z)\right]\\
&= \prod_{i=1}^{n} a_i 2^{n-1}\left[(a_{n+1}+z)\sum_{k=0}^{n}\frac{(-n)_k}{(-2n+1)_k}\frac{z^k}{k!}+a_{n+1}\sum_{k=0}^{n-1}\frac{(-n+1)_k}{(-2n+1)_k}\frac{z^k}{k!}\right].
\end{aligned} \tag{57}$$

Next, we collect the terms of the same degree k in (57) and simplify the obtained coefficients. The coefficients in front of z^k, $k = 2, 3, \ldots, n$,

$$\begin{aligned}
& a_{n+1}\frac{(-n)_k}{(-2n+1)_k}\frac{1}{k!} + a_{n+1}\frac{(-n+1)_k}{(-2n+1)_k}\frac{1}{k!} + \frac{(-n)_{k-1}}{(-2n+1)_{k-1}}\frac{1}{(k-1)!} \\
= \; & a_{n+1}(-n+k-1)\frac{(-n+1)_{k-2}}{(-2n+1)_{k-1}}\frac{1}{k!} + (-n)\frac{(-n+1)_{k-2}}{(-2n+1)_{k-1}}\frac{1}{(k-1)!} \\
= \; & [a_{n+1}(-n+k-1) + (-n)(k)]\frac{(-n+1)_{k-2}}{(-2n+1)_{k-1}}\frac{1}{k!} \\
= \; & [(2n+1)(-n+k-1) + (-n)(k)]\frac{(-n+1)_{k-2}}{(-2n+1)_{k-1}}\frac{1}{k!} \qquad (58) \\
= \; & (-2n+k-1)(n+1)\frac{(-n+1)_{k-2}}{(-2n+1)_{k-1}}\frac{1}{k!} \\
= \; & (-2n+k-1)(n+1)\frac{(-2n-1)(-2n)}{(-n-1)(-n)(-2n+k-1)}\frac{(-n-1)_k}{(-2n-1)_k}\frac{1}{k!} \\
= \; & 2a_{n+1}\frac{(-n-1)_k}{(-2n-1)_k}\frac{1}{k!}.
\end{aligned}$$

For the constant term, $k = 0$, we have

$$a_{n+1} \cdot 1 + a_{n+1} \cdot 1 = 2a_{n+1} \cdot 1. \qquad (59)$$

For the term of degree 1,

$$a_{n+1}\frac{-n}{-2n+1} + 1 + a_{n+1}\frac{-n+1}{-2n+1} = 2(n+1) = 2a_{n+1}\frac{-n-1}{-2n-1}. \qquad (60)$$

For the term of degree $n+1$

$$\frac{(-n)_n}{(-2n+1)_n}\frac{1}{n!} = \frac{(-2n-1)(-2n)}{(-n-1)(-n)}\frac{(-n-1)_{n+1}}{(-2n-1)_{n+1}}\frac{1}{n!} = 2a_{n+1}\frac{(-n-1)_{n+1}}{(-2n-1)_{n+1}}\frac{1}{(n+1)!}. \qquad (61)$$

Inserting (58)-(61) into (57), we obtain

$$Q_{n+1}(z) \;=\; \prod_{i=1}^{n+1} a_i 2^n \sum_{k=0}^{n+1} \frac{(-n-1)_k}{(-2n-1)_k}\frac{z^k}{k!} = \prod_{i=1}^{n+1} a_i 2^n {}_1F_1(-n-1, -2n-1, z). \qquad (62)$$

Statement (56b) can be proven using a similar reasoning. To avoid repetition, the proof is omitted. Thus, we proved that (56) is valid for any $n \in \mathbb{N}$. □

Now we can complete the proof of Theorem 1.

Proof. Let $\mathbf{w}_n = \tilde{A}_n^{-1}\mathbf{r}_n$, i.e. $\tilde{A}_n\mathbf{w}_n = \mathbf{r}_n$. Using the Cramer's rule, $\mathbf{w}_n = \det(\tilde{A}_n)^{-1}[M_n^1, M_n^2, \ldots, M_n^n]^T$. Therefore,

$$\begin{aligned}
f_n(z) \;&=\; [1, 1, \ldots, 1]\tilde{A}_n^{-1}\mathbf{r}_n \\
&=\; [1, 1, \ldots, 1]\mathbf{w}_n \\
&=\; \frac{1}{\det(\tilde{A}_n)}\sum_{j=1}^{n} M_n^j \\
&=\; \frac{R_n(-z)}{Q_n(z)} \qquad \text{(Lemma 1 and Lemma 2)} \\
&=\; \frac{{}_1F_1(-n+1, -2n+1, -z)}{{}_1F_1(-n, -2n+1, z)} \qquad \text{(Lemma 3)}
\end{aligned}$$

□

4 Spectrum of the DG discretization

In the previous section we showed that the eigenvalues of the discontinuous Galerkin spatial discretization matrix L are given by

$$f_n(\lambda) = \omega_j, \quad \omega_j = e^{\frac{2\pi i}{N} j}, \quad j = 0, 1, \dots, N-1. \tag{63}$$

Then, using (40) the eigenvalues can be computed as n roots of

$$R_n(-\lambda) - \omega_j Q_n(\lambda) = 0 \tag{64}$$

for each $0 \leq j \leq N-1$.

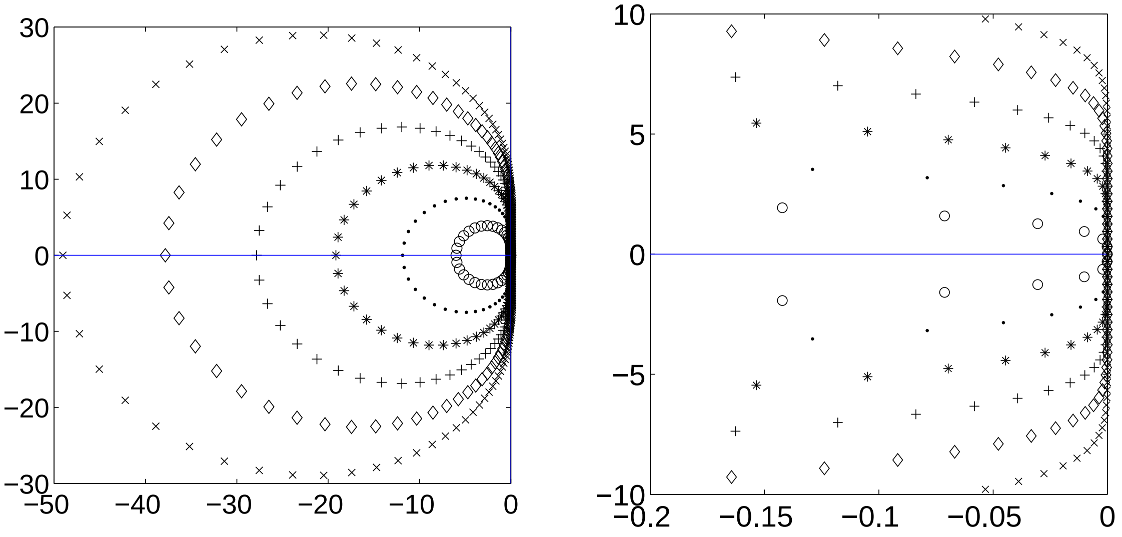

Figure 2: Eigenvalues of L for $p = 1, 2, 3, 4, 5, 6$ with $N = 20$. The inner curve corresponds to $p = 1$ and the outer curve corresponds to $p = 6$. The right figure is a zoom of the left one.

The computed eigenvalues with $p = 1, 2, \dots, 6$ on a twenty cell mesh are shown in Figure 2, left. We observe that the size of the spectrum grows with p. At the same time the curves $|f_{p+1}(z)| = 1$ (of which (63) is a discrete approximation) seem to approach and flatten near the imaginary axis (Fig. 2, right) as p increases.

The leftmost eigenvalues are mainly responsible for the decrease of the CFL number with increasing order of approximation and can be used as a proxy for the size of the spectrum. We investigate the growth rate of the eigenvalue with the largest magnitude later in this section. Flattening of the spectral curves near the imaginary axis means that for the fully discrete method to be stable with a constant (mesh size independent) CFL number, the absolute stability region of the time integration scheme should include a sufficiently large part of the imaginary axis. For example, the DG with $p > 0$ is not stable with a fixed CFL number with the forward Euler time stepping or $p > 1$ and a two stage second order Runge-Kutta schemes [7]. However, despite approaching the imaginary axis, the eigenvalues are never purely imaginary when the upwind flux is used in the DG discretization. This is proven in the following lemma and theorem.

Lemma 4. *Let $Q_n(z)$ and $R_n(z)$ be polynomials defined by (39). Then,*

$$Q_n(\beta i)Q_n(-\beta i) = R_n(\beta i)R_n(-\beta i) + \beta^{2n}, \quad \beta \in \mathbb{R}, \quad n = 1, 2, \ldots. \tag{65}$$

Proof. When $\beta = 0$, it follows from (56) and the definition (36) that $Q_n(0) = R_n(0)$. Then, (65) is trivially true.

When $\beta \neq 0$, we will use the mathematical induction on n to prove (65). For $n = 1$, from (39), we obtain

$$Q_1(\beta i)Q_1(-\beta i) = (1+\beta i)(1-\beta i) = 1+\beta^2 = R_1(\beta i)R_1(-\beta i) + \beta^2, \tag{66}$$

which establishes the basis of induction. We assume (65) is valid for n, and we will prove it consequently holds for $n+1$. Using (39) yields

$$\begin{aligned} Q_{n+1}(\beta i)Q_{n+1}(-\beta i) &= [(a_{n+1}+\beta i)Q_n(\beta i) + a_{n+1}R_n(\beta i)]\,[(a_{n+1}-\beta i)Q_n(-\beta i) + a_{n+1}R_n(-\beta i)] \\ &= (a_{n+1}^2+\beta^2)Q_n(\beta i)Q_n(-\beta i) + a_{n+1}^2 R_n(\beta i)R_n(-\beta i) \\ &\quad +(a_{n+1}^2 + a_{n+1}\beta i)Q_n(\beta i)R_n(-\beta i) + (a_{n+1}^2 - a_{n+1}\beta i)Q_n(-\beta i)R_n(\beta i). \end{aligned} \tag{67}$$

$$\begin{aligned} R_{n+1}(\beta i)R_{n+1}(-\beta i) &= [a_{n+1}Q_n(\beta i) + (a_{n+1}-\beta i)R_n(\beta i)][a_{n+1}Q_n(-\beta i) + (a_{n+1}+\beta i)R_n(-\beta i)] \\ &= a_{n+1}^2 Q_n(\beta i)Q_n(-\beta i) + (a_{n+1}^2+\beta^2)R_n(\beta i)R_n(-\beta i) \\ &\quad +(a_{n+1}^2 + a_{n+1}\beta i)Q_n(\beta i)R_n(-\beta i) + (a_{n+1}^2 - a_{n+1}\beta i)Q_n(-\beta i)R_n(\beta i). \end{aligned} \tag{68}$$

Thus,

$$\begin{aligned} Q_{n+1}(\beta i)Q_{n+1}(-\beta i) - R_{n+1}(\beta i)R_{n+1}(-\beta i) &= \beta^2[Q_n(\beta i)Q_n(-\beta i) - R_n(\beta i)R_n(-\beta i)] \\ &= \beta^{2(n+1)}, \end{aligned} \tag{69}$$

which completes the proof. □

Theorem 2. *Equation (63) has no pure imaginary roots.*

Proof. We start by observing that for any polynomial $p(z)$ with real coefficients the following holds

$$\overline{p(\beta i)} = p(-\beta i). \tag{70}$$

Next, let us assume that $z = \beta i$ is a pure imaginary root of (63), where $\beta \neq 0$ is a real number.

Substitute $z = \beta i$ into (40) and take the modulus to obtain

$$\begin{aligned} |f_n(\beta i)|^2 &= f_n(\beta i)\overline{f_n(\beta i)} \\ &= f_n(\beta i)\, f_n(-\beta i) \\ &= \frac{R_n(\beta i)R_n(-\beta i)}{Q_n(\beta i)Q_n(-\beta i)} \\ &= \frac{R_n(\beta i)R_n(-\beta i)}{R_n(\beta i)R_n(-\beta i)+\beta^{2n}} \quad \text{(Lemma 4)} \\ &= \frac{|R_n(\beta i)|^2}{|R_n(\beta i)|^2+\beta^{2n}} < 1. \end{aligned} \tag{71}$$

Since $|f_n(z)| = 1$ is a necessary condition for z being a root of (63), $|f_n(\beta i)| < 1$ implies that βi is not a root of (63). □

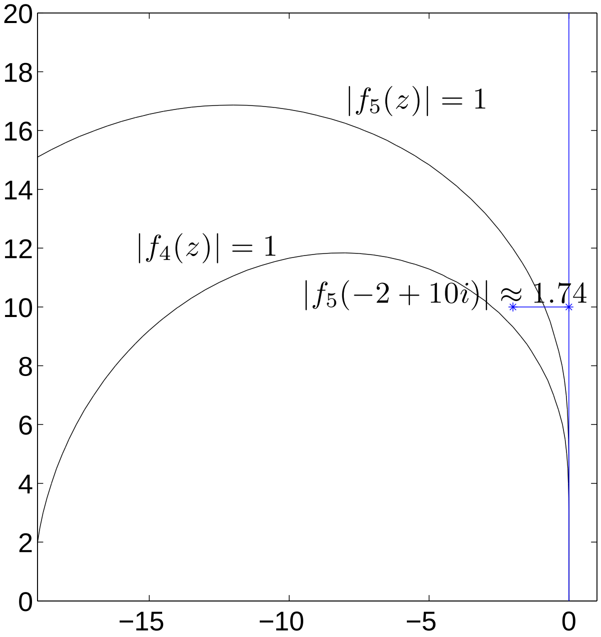

Figure 3: Illustration of $|f_n(z)| = 1$ approaching the imaginary axis. For a randomly chosen point $\hat{z} = -2+10i$, we can find $|f_5(\hat{z})| > 1$, so $|f_5(z)| = 1$ passes through the right of this point $\hat{z}$.

Padé approximants of the exponential function converge to the exponent at every point in the complex plane [11], p.531-536. If we pick an arbitrary point $z = \alpha + \beta i,\ \alpha < 0$, from the left half plane, we have

$$|\lim_{n\to\infty} f_n(z)| = |e^{-z}| > 1, \tag{72}$$

i.e. there exists a sufficiently large N such that $|f_N(z)| > 1$. In the proof of Theorem 2, we showed that $|f_N(\beta i)| < 1$. Assuming that $|f_N(z)|$ is analytic on the line $\mathrm{Im}\, z = \beta$ (since there exists only a finite number of poles of $|f_N(z)|$ this is not a restrictive assumption), there exists a point $z' = \alpha' + \beta i$, where $\alpha < \alpha' < 0$, such that $|f_N(z')| = 1$. This implies that the curve $|f_N(z)| = 1$ goes across the region between $z = \alpha + \beta i$ and βi. Since the point z is randomly chosen from the left half plane, we can conclude that for any point close enough to the imaginary axis, there exists a curve which is even closer to the imaginary axis. Thus, we demonstrated that $|f_n(z)| = 1$ approach the imaginary axis as n grows. The reasoning is illustrated for a particular choice of a point z and $n = 5$ in Figure 3.

Next, we analyze the growth of the eigenvalue largest in modulus. We conjecture it to be the real eigenvalue located on the leftmost part of the spectral curves $|f_n(z)| = 1$ in Figure 2. All roots of (63) are located in the left half of the complex plane. This can be seen from Theorem 4.12 in [10], which states that the curve $|R(z)| = 1$, where $R(z)$ is the $[p/p+1]$ Padé approximant of $exp(z)$ is located in the right half of the complex plane. Since $f_n(z)$ is the same approximant to $exp(-z)$, $|f_n(z)| = 1$ is a mirror image of $|R(z)| = 1$ with respect to the imaginary axis, and the result follows. Below we make a few simple statements about the roots of (63).

Proposition 1. *Equation (63) always has a zero root.*

Proof. ¿From (56) and (36), the zero order coefficients in $R(-z)$ and $Q(z)$ are the same and are equal to $\prod_{i=1}^{n} a_i 2^{n-1}$. Using (40), $f_n(0) = R(0)/Q(0) = 1$. □

Proposition 2. *The real roots of $f_n(z) = \omega_k$ correspond to $\omega_k = \pm 1$.*

Proof. Consider $f_n(z) = \omega_k$ with a real z. Since f_n is a rational function with real coefficients, the right hand side must be a real number. Hence, $\omega_k = \pm 1$. □

Depending on the number of mesh cells N and the order of approximation p, there might be one real root or a couple of complex conjugate numbers on the left most part of the curve. This is not essential as (63) is a discrete version of $|f_n(z)| = 1$. Below we state the conditions for (63) to have a real negative root which we will denote by z^* and without loss of generality we will assume that the mesh is such that it exists.

Proposition 3. *For a discretization with an even number of cells N, (63) has at least one non-zero real root.*

Proof. If n is an odd number, we rewrite $f_n(z) = -1$ as $R(-z) + Q(z) = 0$. Since the zero order coefficient in $R(-z)$ and $Q(z)$ is the same (Proposition 1), zero is not a root of $R(-z) + Q(z) = 0$. Since n is odd, there exists at least one non-zero real root.

If n is an even number, we rewrite $f_n(z) = 1$ as $R(-z) - Q(z) = 0$. Since the zero order coefficient in $R(-z)$ and $Q(z)$ is the same, $R(-z) - Q(z) = 0$ can be expressed as $zr(z) = 0$, where $r(z)$ is a real polynomial of degree $n - 1$. Consequently, it should have one real root which cannot be zero because, as shown in (56) and (36), the first order coefficients of $R(-z)$ and $Q(z)$ are not the same. □

For an odd number of cells N, an odd degree approximation (even n) results in at least one nonzero real root. With an even degree of approximation, we conjecture that the only real root is zero. Since the eigenvalues always locate on $|f_n(z)| = 1$, which does not depend on N, we can assume N is even for analyzing the size of the spectrum.

If the negative real root z^* exists, it is conjectured to have the largest modulus, and this largest modulus also performs as a bound of all roots when z^* does not exist. Next, we will derive a bound on z^*.

Using (38) to write (63) in a polynomial form

$$p(z) = {}_1F_1(-n+1,-2n+1,-z) - e^{\frac{2\pi i}{N}j}{}_1F_1(-n,-2n+1,z), \tag{73}$$

and collecting the terms of the same order, we obtain

$$p(z) = c_n z^n + c_{n-1}z^{n-1} + \cdots + c_1 z + c_0, \tag{74}$$

where

$$\begin{aligned} c_n &= -e^{\frac{2\pi i}{N}j}\tfrac{(-n)_n}{(-2n+1)_n}\tfrac{1}{n!} \\ c_{n-1} &= -e^{\frac{2\pi i}{N}j}\tfrac{(-n)_{n-1}}{(-2n+1)_{n-1}}\tfrac{1}{(n-1)!} + \tfrac{(-n+1)_{n-1}}{(-2n+1)_{n-1}}\tfrac{1}{(n-1)!} \\ &\vdots \\ c_0 &= -e^{\frac{2\pi i}{N}j} + 1. \end{aligned} \tag{75}$$

Since the sum of all the roots of $p(z) = 0$ satisfies

$$\sum_{i=1}^{n} z_i = -\frac{c_{n-1}}{c_n} = -n^2 - ne^{-\frac{2\pi i}{N}j}, \tag{76}$$

we obtain that $\mathrm{Re}(-\frac{c_{n-1}}{c_n}) \geqslant -n(n+1)$. Noticing that all the roots have non-positive real parts, $-n(n+1)$ is a lower bound of the real part of all roots including z^* for all n. We are interested whether this bound is tight and reasonably well represents the growth speed of the largest root. Table 1 lists the roots of the largest modulus up to order 24, and Figure 4 shows the absolute value of these roots up to order 100. We see that the bound overestimates the roots especially for large n. Next, we show that the asymptotic growth rate is not quadratic. In particular, the following theorem proves that $-cn^2$ is not a root of (63) for all $c > 0$.

Theorem 3. *For all $c > 0$, $|f_n(-cn^2)| \propto \frac{1}{n}$ as $n \to \infty$.*

Proof. Consider the hypergeometric function ${}_1F_1(a,b,z)$ defined in (36) and (37). When a and b are negative integers, the function ${}_1F_1(a,b,z)$ is a polynomial of degree $|a|$. We factor out the term of the highest degree of z and define the function

$$G(a,b,z) = \frac{(b)_{|a|}}{(a)_{|a|}}\frac{|a|!}{z^{|a|}}{}_1F_1(a,b,z), \tag{77}$$

or, in an explicit form,

$$G(a,b,z) = \sum_{k=0}^{|a|}\frac{c_k}{z^k}, \quad c_k = C_{|a|}^k (a-b+1)_k, \tag{78}$$

where $C_{|a|}^k$ are binomial coefficients. Substituting (77) into (38) yields

$$f_n(z) = \frac{{}_1F_1(-n+1,-2n+1,-z)}{{}_1F_1(-n,-2n+1,z)} = \frac{(-1)^{n-1}n}{z}\frac{G(-n+1,-2n+1,-z)}{G(-n,-2n+1,z)}. \tag{79}$$

Table 1: Real eigenvalues of L on a two cell grid.

p	1	2	3	4	5	6
$-z^*$	6	11.8424	19.1569	27.8419	37.8247	49.0518
$(p+1)(p+2)$	6	12	20	30	42	56
p	7	8	9	10	11	12
$-z^*$	61.4815	75.0797	89.8181	105.6720	122.6204	140.6442
$(p+1)(p+2)$	72	90	110	132	156	182
p	13	14	15	16	17	18
$-z^*$	159.7268	179.8529	201.0087	223.1817	246.3603	270.5337
$(p+1)(p+2)$	210	240	272	306	342	380
p	19	20	21	22	23	24
$-z^*$	295.6920	321.8258	348.9264	376.9857	405.9960	435.9500
$(p+1)(p+2)$	420	462	506	552	600	650

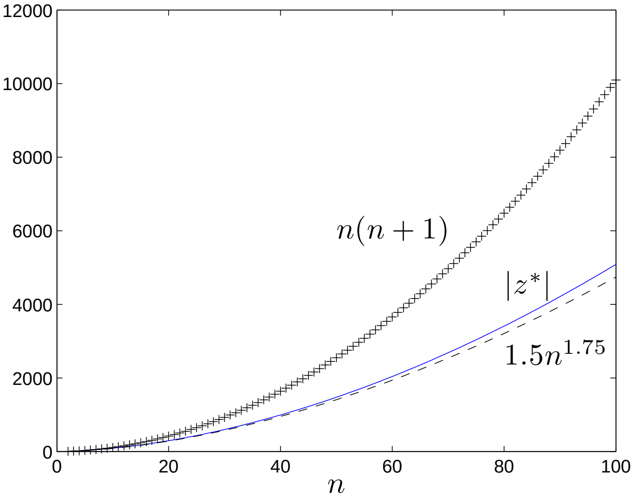

Figure 4: Absolute value of the negative real roots on a two cell grid, the upper bound $n(n+1)$ and the lower bound $1.5n^{1.75}$ as a function of $n = p+1$.

Next, we will prove that $\lim\limits_{n\to\infty} G(-n,-2n+1,-cn^2) = e^{-\frac{1}{c}}$. Substituting $a=-n, b=-2n+$

$1, z = -cn^2$ into (78) gives

$$\begin{aligned}
G(-n,-2n+1,-cn^2) &= \sum_{k=0}^{n} C_n^k \frac{(n)_k}{(-cn^2)^k} \\
&= \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \frac{(n)_k}{(-cn^2)^k} \\
&= \sum_{k=0}^{n} \frac{(-1)^k}{c^k k!} [\frac{n!}{(n-k)!} \frac{(n)_k}{(n^2)^k}] \\
&= 1 + \sum_{k=1}^{n} \frac{(-1)^k}{c^k k!} (1-\frac{1}{n^2})(1-\frac{2^2}{n^2}) \cdots (1-\frac{(k-1)^2}{n^2}).
\end{aligned}$$

For simplicity, we call

$$\tilde{d}_k = \frac{(-1)^k}{c^k k!}, \quad e_0^n = 1, \quad e_k^n = (1-\frac{0}{n^2})(1-\frac{1}{n^2})(1-\frac{2^2}{n^2}) \cdots (1-\frac{(k-1)^2}{n^2}), \quad k = 1,2,..., \tag{80}$$

and define d_k^n as

$$d_k^n = \begin{cases} 1, & k = 0, \\ \frac{(-1)^k}{c^k k!}(1-\frac{0}{n^2})(1-\frac{1}{n^2})(1-\frac{2^2}{n^2}) \cdots (1-\frac{(k-1)^2}{n^2}) = \tilde{d}_k e_k^n, & k = 1,...,n, \\ 0, & k > n. \end{cases} \tag{81}$$

We also define two partial sums

$$D_l^n = \sum_{k=0}^{l} d_k^n, \quad \tilde{D}_l = \sum_{k=0}^{l} \tilde{d}_k, \quad l = 0,1,.... \tag{82}$$

Then, $G(-n,-2n+1,-cn^2) = D_n^n$. Since for a fixed k $\lim_{n\to\infty} e_k^n = 1$, we conclude that

$$\lim_{n\to\infty} d_k^n = \tilde{d}_k, \quad \forall k \geqslant 0. \tag{83}$$

Noticing that $\lim_{l\to\infty} \tilde{D}_l = e^{-\frac{1}{c}}$, we have

$$\forall\, \varepsilon > 0, \exists\, K_1 > 0, s.t. \forall\, k > K_1, \left| \tilde{D}_{K_1} - e^{-\frac{1}{c}} \right| < \varepsilon. \tag{84}$$

And from the definition of $\tilde{d}_k$,

$$\exists\, K_2 > 0, s.t. \forall\, k > K_2, |\tilde{d}_{K_2+1}| < \varepsilon. \tag{85}$$

Let $K = \max(K_1, K_2)$, then

$$\lim_{n\to\infty} D_K^n = \sum_{k=0}^{K} \lim_{n\to\infty} d_k^n = \sum_{k=0}^{K} \tilde{d}_k = \tilde{D}_K. \tag{86}$$

The expression above implies that

$$\exists N > K > 0,\ s.t. \forall\, n > N,\ |D_K^n - \tilde{D}_K| < \varepsilon. \tag{87}$$

On the other hand, since $\{d_k^n\}$ have alternating signs while $|d_k^n|$ decrease as $k \to \infty$,

$$|D_K^n - D_n^n| < |d_{K+1}^n| < |\tilde{d}_{K+1}| < \varepsilon.\ \forall\, n > K \tag{88}$$

Combining (84), (87) and (88), we obtain

$$|D_n^n - e^{-\frac{1}{c}}| \leqslant |D_n^n - D_K^n| + |D_K^n - \tilde{D}_K| + |\tilde{D}_K - e^{-\frac{1}{c}}| < 3\varepsilon, \quad n > N,$$

i.e.

$$\lim_{n\to\infty} G(-n, -2n+1, -cn^2) = \lim_{n\to\infty} D_n^n = e^{-\frac{1}{c}}. \tag{89}$$

Using the same reasoning, we can prove that

$$\lim_{n\to\infty} G(-n+1, -2n+1, cn^2) = e^{\frac{1}{c}}. \tag{90}$$

Combining (89), (90) and (79) yields

$$\lim_{n\to\infty} |f_n(-cn^2)| = \lim_{n\to\infty} \left| \frac{G(-n+1, -2n+1, cn^2)}{G(-n, -2n+1, -cn^2)} \right| \frac{n}{cn^2} = \lim_{n\to\infty} e^{\frac{2}{c}} \frac{1}{cn} = 0. \tag{91}$$

□

We have proved that regardless of the constant c, $|f_n(cn^2)|$ is small for large enough n and consequently cannot be a root of (63). In other words, any quadratic function will overcome the curve $|z^*(n)|$ (Fig. 4). If we assume that the real root z^* grows as a power function $-cn^\alpha$, then for $\alpha > 2$, by following the steps in the proof of Theorem 2 we can show that $\lim_{n\to\infty} G(-n, -2n+1, -cn^\alpha) = 1$ and $\lim_{n\to\infty} G(-n+1, -2n+1, cn^\alpha) = 1$. So, in this case, $\lim_{n\to\infty} f_n(-cn^\alpha) = 0$ also implies $z^* = -cn^\alpha$ is not the proper estimate for the root. We conclude that the spectrum of L should grow slower than $-cn^\alpha$. Numerical experiments reveal that $-1.5n^{1.75}$ is an upper bound on z^* for all n (Fig. 4). Least square fitting $-cn^\alpha$ for the first one hundred roots gives $-1.4n^{1.78}$. Bounds on the eigenvalues for very large n are reported in Figure 5. The computations were performed for a two cell grid using MATLAB. They are believed to be accurate in the sense of small error in $f_n(z^*) - 1$.

Remark. We should mention that the Padé approximants $f_n(z)$ are only a good approximation to e^{-z} in regions close to the origin of the complex plane. In Figure 6 we plot $|f_{p+1}(z)|$ for real negative z with $p = 1, 2, 3$. The spike in the $p = 2$ plot is related to the nearby pole of $f_3(z)$. We note that this behavior, i.e. that the exponential function grows in magnitude while the Padé approximants decay to zero as $|z|$ increases, is similar for complex z but is more difficult to illustrate. Comparing Figures 2 and 6 reveals that for many eigenvalues λ, $|f_n(\lambda)|$ is far from $e^{-\lambda}$. Although the region where $f_n(z) \approx e^{-z}$ grows with n, it grows slower than the eigenvalues. Consequently, the approximant f_n should not be viewed as an approximation of $exp(-z)$ as far as eigenvalues of L are concerned.

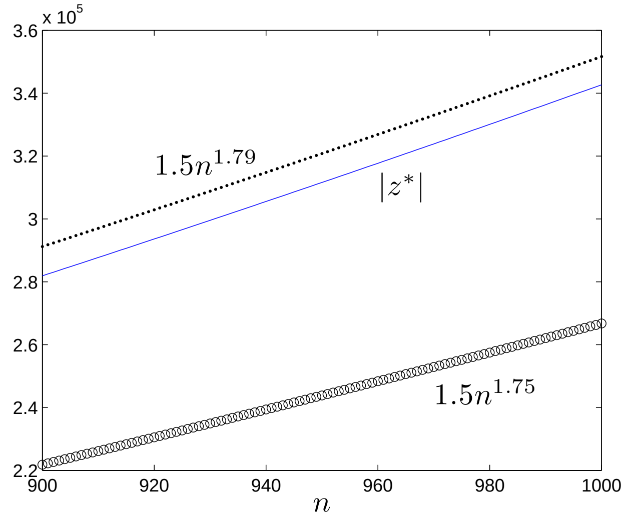

Figure 5: Absolute value of the negative real roots on a two cell grid and two bounds for large values of n.

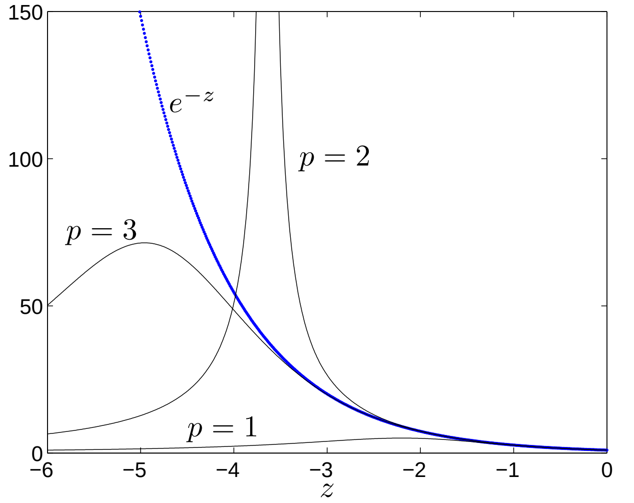

Figure 6: Comparison of e^{-z} with $|f_{p+1}(z)|$ for $p = 1, 2, 3$.

5 Conclusions and future work

We have derived a closed form expressions for the eigenvalues of the DG spatial discretization applied to the one-dimensional linear advection equation with periodic boundary conditions and the upwind flux. We have proven that the characteristic polynomial of the spatial discretization matrix L is related to the subdiagonal $[p/p+1]$ Padé approximant of e^{-z}. Based on the analytical equation for the eigenvalues, we have proven that there is no pure imaginary eigenvalues. We have also shown that $(p+1)(p+2)$ is a guaranteed bound on the size of the eigenvalues which can be used to compute the CFL condition for large p. However, we have also proven that the growth rate of the largest eigenvalue is less than $(p+1)^2$. We conjecture that a more accurate rate is proportional to $(p+1)^{1.75}$. This is in contrast with the currently assumed quadratic rate for the DGM [13] and various spectral methods [12]. A more accurate analytical estimate would be of interest.

A potential application would be to use this result to improve the CFL number of the DGM and, consequently, its computational efficiency. We show [4], that the coefficients of the scheme can be manipulated to decrease the radius of the spectrum, i.e. to increase the CFL number, while preserving the convergence rate in the L^2 norm. The improvement depends on the order of approximation. For example, we can have an improvement up to a factor of three for $p = 1$ and up to a factor of 5.5 for $p = 3$. We also apply this result to analysis of the spectrum on non-uniform grids [15]. In particular, we are interested how a global CFL condition is affected by the composition of the mesh and how a few small cells influence the overall stability of the method. The improvement is observed in some special cases such as Cartesian grids with embedded geometries.

6 Acknowledgment

This research was supported in part by Natural Sciences and Engineering Research Council (NSERC) of Canada grant 341373-07

Appendices

Table 2: Part of the Padé table of e^z [3]

	0	1	2	3
0	$\frac{1}{1}$	$\frac{1+z}{1}$	$\frac{2+2z+z^2}{2}$	$\frac{6+6z+3z^2+z^3}{6}$
1	$\frac{1}{1-z}$	$\frac{2+z}{2-z}$	$\frac{6+4z+z^2}{6-2z}$	$\frac{24+18z+6z^2+z^3}{24-6z}$
2	$\frac{2}{2-2z+z^2}$	$\frac{6+2z}{6-4z+z^2}$	$\frac{12+6z+z^2}{12-6z+z^2}$	$\frac{60+36z+9z^2+z^3}{60-24z+3z^2}$
3	$\frac{6}{6-6z+3z^2-z^3}$	$\frac{24+6z}{24-18z+6z^2-z^3}$	$\frac{60+24z+3z^2}{60-36z+9z^2-z^3}$	$\frac{120+60z+12z^2+z^3}{120-60z+12z^2-z^3}$

Table 3: Polynomials $Q_n(z)$ and $R_n(-z)$ defined in (39)

n	$R_n(-z)$	$Q_n(z)$
2	$6(1-\frac{1}{3}z)$	$6(1+\frac{2}{3}z+\frac{1}{6}z^2)$
3	$60(1-\frac{2}{5}z+\frac{1}{20}z^2)$	$60(1+\frac{3}{5}z+\frac{3}{20}z^2+\frac{1}{60}z^3)$
4	$840(1-\frac{3}{7}z+\frac{1}{14}z^2-\frac{1}{210}z^3)$	$840(1+\frac{4}{7}z+\frac{1}{7}z^2+\frac{2}{105}z^3+\frac{1}{840}z^4)$
5	$15120(1-\frac{4}{9}z+\frac{1}{12}z^2-\frac{1}{126}z^3+\frac{1}{3024}z^4)$	$15120(1+\frac{5}{9}z+\frac{5}{36}z^2+\frac{5}{252}z^3+\frac{5}{3024}z^4+\frac{1}{15120}z^5)$

.

References

[1] M. Abramowitz and I.A. Stegun, editors. *Handbook of Mathematical Functions*. Dover, New York, 1965.

[2] M. Ainsworth. Dispersive and dissipative behaviour of high order discontinuous Galerkin finite element methods. *Journal of Computational Physics*, 198:106–130, 2004.

[3] G. A. Baker and P. R. Graves-Morris. *Padé Approximants*. Addison-Wesley, Reading, Mass.; Don Mills, Ont., 1981.

[4] N. Chalmers, L. Krivodonova, and R. Qin. Relaxing the CFL number of the Discontinuous Galerkin method. In preparation.

[5] G. Chavent and B. Cockburn. The local projection P^0P^1 discontinuous Galerkin method for scalar conservation laws. *RAIRO, Model. Math. Anal. Numer.*, 23:565, 1989.

[6] B. Cockburn and C.-W. Shu. The Runge-Kutte local projection P^1 discontinuous Galerkin method for scalar conservation laws. *RAIRO Model. Math. Anal. Numer.*, 25:337–361, 1991.

[7] B. Cockburn and C.-W. Shu. Runge-Kutta discontinuous Galerkin methods for convection-dominated problems. *Journal of Scientific Computing*, 16:173–261, 2001.

[8] N. Gödel, S. Schomann, T. Warburton, and M. Clemens. GPU accelerated Adams-Bashforth multirate discontinuous Galerkin simulation of high frequency electromagnetic fields. *IEEE Transactions on magnetics*, 48(8):2735–2738, 2010.

[9] S. Gottlieb, D. Ketcheson, and C.-W. Shu. High-order stability preserving time discretizations. *Journal of Scientific Computing*, 38(3):251, 2009.

[10] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II. Stiff and differential-algebraic problems*. Springer, Berlin, second edition, 2002.

[11] P. Henrici. *Applied and Computational Complex Analysis. Special functions-integral transforms-asymptotics-continued fractions*, volume 2. John Wiley & Sons, 1977.

[12] J. S. Hesthaven, S. Gottlieb, and D. Gottlieb. *Spectral Methods for time dependent problems*. Cambridge University Press, Cambridge, 2007.

[13] J. S. Hesthaven and T. Warburton. *Nodal Discontinuous Galerkin Methods. Algorithms, Analysis, and Applications*. Springer, 2007.

[14] F. Q. Hu and H. L. Atkins. Eigensolution analysis of the discontinuous Galerkin method with nonuniform grids. *Journal of Computational Physics*, 182:516–545, 2002.

[15] L. Krivodonova and R. Qin. Linear stability analysis of the discontinuous Galerkin method on non-uniform grids. In preparation.

[16] E. J. Kubatko, C. Dawson, and J. J. Westerink. Time step restrictions for Runge-Kutta discontinuous Galerkin methods on triangular grids. *Journal of Computational Physics*, 227:9697–9710, 2008.

[17] P. Le Saint and P. Raviart. On a finite element method for solving the neutron transport equation. In C. de Boor, editor, *Mathematical Aspects of Finite Elements in Partial Differential Equations*, pages 89–145, New York, 1974. Academic Press.