Relaxing the CFL Number of the Discontinuous Galerkin Method

N. Chalmers * L. Krivodonova * R. Qin†

June 5, 2014

Abstract

We propose a family of high order methods for the solution of hyperbolic conservation laws which are based on the discontinuous Galerkin (DG) spatial discretization. In the standard DG method, the dispersion and dissipation errors and the spectrum of the semi-discrete scheme are related to the $[\frac{p}{p+1}]$ Padé approximants of $\exp(z)$ and $\exp(-z)$. These Padé approximants are responsible for the superconvergent $\mathcal{O}(h^{2p+2})$ and $\mathcal{O}(h^{2p+1})$ errors in dispersion and dissipation, respectively, and the restriction of the CFL number when increasing the order of approximation, p. By modifying the DGM we obtain different rational approximations of the exponential, thereby sacrificing some of the superconvergence of the method, and construct new schemes which allow larger time steps than the original DGM, while having the same order of convergence in the $\mathcal{L}^2$ norm. This is achieved through modifications to the numerical flux. The schemes preserve the attractive properties of the usual DGM, such as the high order accuracy and compact stencil.

1 Introduction

The discontinuous Galerkin (DG) spatial discretization applied to convection problems has maximum a Courant-Friedrichs-Lewy (CFL) number that decreases with the order of approximation p as (approximately) $1/(2p+1)$ when paired with an appropriate order explicit Runge-Kutta scheme. This rather restrictive condition is caused by the growth of the spectrum of the spatial discretization operator of the semi-discrete scheme, which increases slightly slower than $\mathcal{O}(p^2)$ [11]. In contrast, finite difference schemes have a stability restriction that grows with the size of the computational stencil as $\mathcal{O}(p)$. This makes the DGM a more expensive scheme for the same theoretical order of convergence. This is often quoted as one of the shortcomings of the DGM. A possible solution to this issue was proposed by Warburton and Hagstrom in [18], in which the authors propose the use of a co-volume mesh which allows an order independent CFL number. However, this method is limited to structured grids and requires mappings of the solution between the original and co-volume meshes. The method in [18] shrinks the spectrum of the DG method so that it does not require the usual $1/(2p+1)$ scaling. Another approach is to devise explicit time-integrators with larger absolute stability regions or stability regions which better encapsulate the spectrum of the DG spatial operator [13, 14, 17]. For Runge-Kutta methods this usually comes at the cost of additional stages.

For the same theoretical order of convergence, numerical schemes can have distinctly different global accuracy. It has been pointed out that the discontinuous Galerkin scheme is more accurate than the finite volume scheme, e.g. when applied to the two-dimensional Euler equations [12], in terms of the $\mathcal{L}^2$ norm. One reason for this is the small dispersive and dissipative errors in the DGM. It was shown [7, 2] that the local dispersion and dissipation errors of the method are $\mathcal{O}(h^{2p+3})$ and $\mathcal{O}(h^{2p+2})$, respectively, for resolved wave numbers. Globally, these errors are each one order lower. These small errors lead to slower accumulation of the numerical error which is especially noticeable for long time calculations.

The accuracy of the DGM, in terms of the dispersive and dissipative errors, has been studied by several authors [8, 15]. Notably, by assuming that the exact solution of the linear advection equation is of the form $\exp(i(kx-\omega t))$, and fixing ω, Hu and Atkins [7] numerically showed that the numerical wave speed k_h of the DG method is related to the exact wave speed by the dispersion relation $f_p(ihk) = \exp(ihk_h)$, where $f_p(z)$ is the sub-diagonal $[\frac{p}{p+1}]$ Padé approximant [3] of $\exp(z)$. Ainsworth gave a more in depth analysis in [2], including explicit expressions for the leading term of the errors. More recently, Krivodonova and Qin [11] showed that the spectrum of the DGM is given

*Department of Applied Mathematics, University of Waterloo, 200 University Avenue West, Waterloo, Ontario Canada, N2L 3G1
†School of Mathematics and Computer Science, Guizhou Normal University, Guiyang, China, 550001.

by the same Padé approximant, but of $\exp(-z)$. In particular, they proved that the spectrum is given by $|f_p| = 1$ and pointed out that the large spectrum and the resulting small CFL number of the DGM are a direct consequence of the superconvergence property. It is reasonable to assume then that a scheme resulting in a different rational approximation of $\exp(z)$ may have desirable properties, e.g. a less restrictive CFL number. The difficulty is to modify the weak DG form to obtain such a scheme.

In this paper, we propose modifications to the DG method which involve $p+1$ parameters $\alpha_m,\ m = 0, 1, \ldots, p$, which we call flux multipliers. In the case when $\alpha_m = 1$, for $k = 0, 1, \ldots, p$, we recover the original DGM. When a certain α_m is not equal to one we refer to this multiplier as 'modified'. In each equation evolving the m-th degree of freedom on element I_j, c_{jm}, in time (see (10) and (11)), we use the flux multiplier α_m to scale the contribution from the jumps in the numerical flux at cell interfaces to the propagation of c_{jm}. The justification of this operation is that the weak DG formulation consists of integrals over cell volumes plus contributions from jumps in the numerical flux at the cell boundaries. For solutions which belong to the finite element space, the flux jumps are equal to zero and, thus, the proposed modifications will not influence the solution accuracy. More generally, they will not affect the formal results on accuracy and convergence originally established by Cockburn and Shu [5, 4], as long as the equation corresponding to the c_{j0} coefficient (i.e., the one corresponding to the constant basis function) is unchanged. We show that the modifications will affect the eigenvalues of the spatial operator of the semi-discrete scheme, and hence, the CFL number.

In order to relax the time step restriction of the standard DG formulation, we search for a set of flux multipliers α_m that provides the largest increase in the CFL number when using the Legendre polynomial basis. The values for any other polynomial basis could be obtain from the presented ones by a simple transformation. In order to compute this set of values, we use linear algebra software to search for α_m so that the size of the spectrum of the modified scheme is smaller than that of the original DGM. We find that for the orders of approximation considered in this work, the CFL number can be improved by a factor of two or more by modifying only the highest multiplier to be $\alpha_p \approx 0.4$. The modification of more than the highest multiplier generally leads to a larger improvement in the CFL for particular combinations of α_m. Using an energy argument we prove that when only the highest multiplier , α_p, is modified the semi-discrete scheme is linearly stable. In this case small modifications to α_p influence only the size of the spectrum. In a general case where more than one multiplier is modified, a particular choice of multipliers can result in an unstable semi-discrete scheme. However, we are able to numerically find a combination of multipliers which results in stable semi-discrete schemes.

Next, we analyse the accuracy of modified schemes. We prove that modifying m highest order multipliers lowers the order of accuracy in dispersion and dissipation by m orders to $\mathcal{O}(h^{2p+2-m})$ and $\mathcal{O}(h^{2p+1-m})$, respectively. Nevertheless, the order of convergence of the scheme in the $\mathcal{L}^1$ norm remains the same regardless of the number of multipliers changed, as long as α_0 remains equal to one. This follows from the standard DG analysis [5, 4], and our numerical experiments. However, we observe in numerical experiments that the magnitude of the global $\mathcal{L}^1$ error increases due to larger dissipative and dispersive errors. In particular, setting a larger number of multipliers to be not equal to one leads to a larger global error.

The proposed schemes can be viewed from a different perspective. Instead of comparing the schemes based on the size of spatial discretization, we can compare them based on the computational effort. That is, instead of increasing the time step size for a fixed mesh, we can fix the time step and proportionally increase the number of cells. We show that with the modified DG scheme, the solution for the same computational effort is noticeably more accurate in terms of the global error. This is especially advantageous for problem which have high frequency waves or fine structures.

The remainder of this paper is organized as follows: In Section 2 we will introduce the discontinuous Galerkin finite element method, and show how it is modified through the introduction of the flux multipliers α_m. We will then prove several results concerning the effects of these multipliers on the accuracy of the DG scheme by using the linear advection equation as a model problem. We will then investigate the stability of the modified scheme and show that we are able to ameliorate the usual stability restriction of the classical DG scheme through suitable choices in the multipliers α_m. We will conclude by showing that the modified scheme preserves the usual order of convergence in the $\mathcal{L}^1$ norm, and we will show how the scheme performs on several test examples including the linear advection equation and the Euler equations. We also give examples where the accuracies of the DG and modified DG schemes are compared on different sized meshes, but equal computation times.

2 Modified Discontinuous Galerkin Discretization

We consider the one-dimensional scalar conservation law

$$u_t + f(u)_x = 0 \tag{1}$$

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

$$\frac{d}{dt}\int_{x_j}^{x_{j+1}} U_j V\, dx + \int_{x_j}^{x_{j+1}} f(U_j)_x V\, dx = 0, \quad \forall V \in \mathcal{P}_p. \tag{2}$$

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

$$x(\xi) = \frac{x_j + x_{j+1}}{2} + \frac{h_j}{2}\xi \tag{3}$$

yields

$$\frac{h_j}{2}\frac{d}{dt}\int_{-1}^{1} U_j V\, d\xi + \int_{-1}^{1} f(U_j)_\xi V\, d\xi = 0, \quad \forall V \in \mathcal{P}_p. \tag{4}$$

At this point, integration by parts is usually performed on the second term in order to express this integral in terms of the contributions from the cell interfaces plus an integral over the interior of the cell. Instead, notice that as a distribution $f(U_j)_x$ is defined as

$$f(U_j)_x = \begin{cases} \left(f(U_j(x_j)) - f(U_j^*)\right)\delta_{x_j}, & x = x_j \\ f(U_j)_x, & x \in (x_j, x_{j+1}) \\ \left(f(U_{j+1}^*) - f(U_j(x_{j+1}))\right)\delta_{x_{j+1}}, & x = x_{j+1} \end{cases}$$

where δ_{x_j} is the Dirac delta function at $x = x_j$, and U_{j+1}^* is the Riemann state at the interface between the j-th and $(j+1)$-th cells. Moreover, the derivative on the interior term is defined classically since U_j is smooth inside I_j. Using this expression for $f(U_j)$ and assuming the test function $V(\xi)$ is continuous across the cell interface, we can write (4) as

$$\frac{h_j}{2}\frac{d}{dt}\int_{-1}^{1} U_j V\, d\xi + \left[f(U_{j+1}^*) - f(U_j(1))\right] V(1) + \left[f(U_j(-1)) - f(U_j^*)\right] V(-1) + \int_{-1}^{1} f(U_j)_\xi V\, d\xi = 0, \quad \forall V \in \mathcal{P}_p. \tag{5}$$

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

$$\int_{-1}^{1} P_m P_i\, d\xi = \frac{2}{2m+1}\delta_{mi}, \tag{6}$$

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]

$$P_m(1) = 1, \qquad P_m(-1) = (-1)^m. \tag{7}$$

The numerical solution can be written in terms of this 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 $V = P_m$, $m = 0, 1, \ldots, p$, and using (7) and (6) results in

$$\frac{h_j}{2m+1}\dot{c}_{jm} = -\left[f(U_{j+1}^*) - f(U_j(1))\right] - (-1)^m\left[f(U_j(-1)) - f(U_j^*)\right] - \int_{-1}^{1} f(U_j)_\xi P_m \, d\xi, \quad m = 0, 1, \ldots, p, \tag{9}$$

where the dot in $\dot{c}_{jm}$ represents differentiation with respect to t.

Notice that the only contributions from neighbouring cells are concentrated in the two jump terms on the right hand side of (9), while the integral term is purely local to the cell I_j. Moreover, when the exact solution of (1) belongs to the finite element space, these two jump terms will be equal to zero. Consequently, modifying these terms will not affect the accuracy of the solution. This motivates us to consider a modified version of (9),

$$\frac{h_j}{2m+1}\dot{c}_{jm} = -\alpha_m\left[f(U_{j+1}^*) - f(U_j(1))\right] - (-1)^m\alpha_m\left[f(U_j(-1)) - f(U_j^*)\right] - \int_{-1}^{1} f(U_j)_\xi P_m \, d\xi, \quad m = 0, 1, \ldots, p. \tag{10}$$

Here we have introduced the parameters $\alpha_m, m = 0, \ldots, p$, which scale the contributions of the flux discontinuities at the cell interfaces to the propagation of the solution coefficients. Note that when $\alpha_m = 1, \forall m$, we recover the original DG scheme. Since many applications of the DG method use a slightly different formulation than (9), in which the integral term of the flux is integrated by parts, we present an alternative form of (10) where we have performed this extra step and rearranged to obtain

$$\frac{h_j}{2m+1}\dot{c}_{jm} = -\alpha_m\left(f(U_{j+1}^*) - (-1)^m f(U_j^*)\right) - (1-\alpha_m)\left(f(U_j(1)) - (-1)^m f(U_j(-1))\right) + \int_{-1}^{1} f(U_j)P_m' \, d\xi, \quad m = 0, 1, \ldots, p. \tag{11}$$

We expect that this scheme, which we will refer to as the modified DG (mDG) scheme, will perform similarly to the original DG on smooth solutions, where the altered jump contributions are small. In the remainder of the paper we will be interested in establishing what effect these parameters will have on the numerical scheme. Since this analysis is difficult to perform on the general formulation, we will instead consider a simple problem: the linear advection equation.

2.1 Linear Advection Equation

Here we are interested in the problem

$$u_t + au_x = 0,$$

where $a > 0$ and constant, and with periodic boundary conditions. Applying the modified DG discretization to this problem, and using the upwind flux so that $U_{j+1}^* = U_j(1)$, (10) reads

$$\frac{h_j}{2m+1}\dot{c}_{jm} = -(-1)^m\alpha_m a\left[U_j(-1) - U_{j-1}(1)\right] - a\int_{-1}^{1}(U_j)_\xi P_m \, d\xi.$$

Substituting (8) into the above expression and using (7) for the boundary terms we obtain

$$\frac{h_j}{2m+1}\dot{c}_{jm} = -(-1)^m\alpha_m a\left[\sum_{i=0}^{p}(-1)^i c_{ji} - \sum_{i=0}^{p} c_{j-1,i}\right] - a\int_{-1}^{1}\left(\sum_{i=0}^{p} c_{ji}P_i'\right)P_m \, d\xi.$$

Collecting common terms of the right hand side results in

$$\dot{c}_{jm} = a\frac{2m+1}{h_j}\left[(-1)^m\alpha_m\sum_{i=0}^{p} c_{j-1,i} - \sum_{i=0}^{p}\left(\int_{-1}^{1} P_i'P_m \, d\xi + (-1)^{m+i}\alpha_m\right)c_{ji}\right]. \tag{12}$$

This can be written in a vector form as

$$\dot{c}_{jm} = a\frac{2m+1}{h_j}\Bigg((-1)^m\alpha_m[1,1,...,1]\mathbf{c}_{j-1} - \left[\int_{-1}^{1} P_0' P_m d\xi + (-1)^m\alpha_m, ..., \int_{-1}^{1} P_p' P_m d\xi + (-1)^{m+p}\alpha_m\right]\mathbf{c}_j\Bigg), \quad (13)$$

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 $\mathbf{c} = [\mathbf{c}_0^T, \mathbf{c}_1^T, \ldots, \mathbf{c}_N^T]^T$, and assuming a uniform grid $h_j = h, \forall j$, equation (13) can be written as

$$\dot{\mathbf{c}} = \frac{a}{h}\mathbf{L}\mathbf{c}. \quad (14)$$

With the assumed periodic boundary conditions, $\mathbf{L}$ is a block matrix of the form

$$\mathbf{L} = \begin{bmatrix} \tilde{A}_p & 0 & 0 & \ldots & 0 & 0 & \tilde{D}_p \\ \tilde{D}_p & \tilde{A}_p & 0 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & \tilde{D}_p & \tilde{A}_p \end{bmatrix}, \quad (15)$$

where $\tilde{D}_p$ and $\tilde{A}_p$ are $(p+1)\times(p+1)$ matrices,

$$\tilde{D}_p = \begin{bmatrix} \alpha_0 & \ldots & \alpha_0 \\ -3\alpha_1 & \ldots & -3\alpha_1 \\ \vdots & & \vdots \\ (-1)^p(2p+1)\alpha_p & \ldots & (-1)^p(2p+1)\alpha_p \end{bmatrix}, \quad (16)$$

$$\tilde{A}_p = -\begin{bmatrix} b_{00}+\alpha_0 & \ldots & b_{p0}+(-1)^p\alpha_0 \\ 3\left(b_{01}-\alpha_1\right) & \ldots & 3\left(b_{p1}+(-1)^{p+1}\alpha_1\right) \\ \vdots & & \vdots \\ (2p+1)\left(b_{0p}+(-1)^p\alpha_p\right) & \ldots & (2p+1)\left(b_{pp}+\alpha_p\right) \end{bmatrix}, \quad (17)$$

where

$$b_{im} = \int_{-1}^{1} P_i' P_m d\xi. \quad (18)$$

We use the '$\sim$' notation on the matrices $\tilde{A}_p$ and $\tilde{D}_p$ to indicate their dependence on the α_m flux multipliers. In what follows, we will drop the '$\sim$' to indicate that $\alpha_m = 1, \forall m$. For a more concise expression for b_{im}, we notice that the derivatives of the Legendre polynomials satisfy [1]

$$(2m+1)P_m = P_{m+1}' - P_{m-1}'. \quad (19)$$

We then derive

$$P_{m+1}' = (2m+1)P_m + (2(m-2)+1)P_{m-2} + (2(m-4)+1)P_{m-4} + \ldots. \quad (20)$$

Using (20) with the orthogonality property of the Legendre polynomials (6) we obtain

$$b_{im} = \int_{-1}^{1} P_i' P_m d\xi = \begin{cases} 0, & i \leqslant m, \\ 1-(-1)^{i-m}, & i > m. \end{cases} \quad (21)$$

3 Accuracy of the mDG Method

Let us investigate the accuracy of the modified DG scheme by examining how accurately the scheme approximates solutions of the form $u(x,t) = e^{i(kx-\omega t)}$. For the exact solution of the linear advection equation the wave number k will be related to the frequency, ω, by $k = \frac{\omega}{a}$. For a specific fixed frequency ω the numerical scheme will produce

an approximate solution of the form $U(x,t) = e^{i(k_h x - \omega t)}$, where k_h is referred to as the numerical wave number. Therefore, the degree to which the scheme produces a numerical wave number which agrees with the exact wave number will give us a measure of the accuracy of the scheme. To investigate this relation between k_h and k, let us assume the numerical solution has the form

$$\mathbf{c}_j(t) = e^{i(jK_h - \omega t)}\hat{\mathbf{c}},$$

where $\hat{\mathbf{c}}$ is a constant vector of the coefficients, and $K_h = k_h h$ is the non-dimensional numerical wave number. Substituting this into (14), and taking into account (15), yields

$$\left(-\frac{i\omega h}{a}I - \tilde{A}_p - \frac{1}{\lambda}\tilde{D}_p\right)\hat{\mathbf{c}} = 0, \tag{22}$$

where $\lambda = e^{iK_h}$. For a non-trivial solution of (22) to exist, the determinant of the matrix in the brackets must be zero. This condition will yield a relation between the non-dimensional numerical wave number K_h, and the non-dimensional exact wave number $K = \frac{\omega h}{a}$. In [11], Krivodonova and Qin investigate this condition for the regular DG scheme ($\alpha_m = 1, \forall m$). We recall their main results in the following theorems.

Theorem 1 (Krivodonova, Qin [11]). *The condition* $\det(-iKI - A_p - \frac{1}{\lambda}D_p) = 0$ *can be written as*

$$\lambda = f_p(iK). \tag{23}$$

Here, f_p *has the form*

$$f_p(z) = \frac{R_p(z)}{Q_p(-z)},$$

where $R_p(z)$ *is a polynomial of degree* p *and* $Q_p(z)$ *is a polynomial of degree* $p+1$*. Furthermore,* $R_p(z)$ *and* $Q_p(z)$ *can be generated through the recursive relations,*

$$R_p(z) = (2p+1)(R_{p-1}(z) + Q_{p-1}(z)) - zR_{p-1}(z),$$
$$Q_p(z) = (2p+1)(R_{p-1}(z) + Q_{p-1}(z)) + zQ_{p-1}(z),$$

together with the initial conditions $R_0(z) = 1$*, and* $Q_0(z) = 1 + z$*.*

Theorem 2 (Krivodonova, Qin [11]). *The function* $f_p(z)$ *is the* $[\frac{p}{p+1}]$ *Padé approximant of* e^z*, i.e.*

$$f_p(z) = e^z + \mathcal{O}(z^{2p+2}), \tag{24}$$

Using (23) and (24) together with $\lambda = e^{iK_h}$ we immediately obtain

$$e^{iK_h} = e^{iK} + \mathcal{O}((iK)^{2p+2}),$$

from which we get the order estimate

$$K_h = K + iC_1K^{2p+2} + C_2K^{2p+3} + \dots,$$

where C_1 and C_2 are constant and real. Using that $K_h = hk_h$ and $K = hk$ we obtain

$$k_h = k + iC_1k^{2p+2}h^{2p+1} + C_2k^{2p+3}h^{2p+2} + \dots. \tag{25}$$

Recall that the numerical solution is of the form $\mathbf{c}_j(t) = e^{i(jhk_h - \omega t)}\hat{\mathbf{c}}$. Hence, the imaginary part of the error in k_h will cause an error in the amplitude of the numerical solution. We therefore call the imaginary part of the error the *dissipation* error. On the other hand, the real part of the error in k_h will cause a shift in the wave number of the numerical solution. We call this error the *dispersion* error. Therefore we see from (25) that the order of the dispersion error of the DG scheme is $\mathcal{O}(h^{2p+2})$ and the order of the dissipation error is $\mathcal{O}(h^{2p+1})$, for the resolved wave numbers. See [7] and [2] for a more complete discussion of the order of accuracy of the DG scheme.

Let us propose an analogous result for the modified DG scheme.

Proposition 1. *The condition* $\det(-iKI - \tilde{A}_p - \frac{1}{\lambda}\tilde{D}_p) = 0$ *can be written as*

$$\lambda = \tilde{f}_p(iK).$$

Here, $\tilde{f}_p$ *has the form*

$$\tilde{f}_p(z) = \frac{\tilde{R}_p(z)}{\tilde{Q}_p(-z)},$$

where $\tilde{R}_p(z)$ *is a polynomial of degree* p *and* $\tilde{Q}_p(z)$ *is a polynomial of degree* $p+1$*. Furthermore,* $\tilde{R}_p(z)$ *and* $\tilde{Q}_p(z)$ *can be generated through the recursive relations*

$$\begin{aligned}
\tilde{R}_p(z) &= (2p+1)\alpha_p q_{p-1}(z) - z\tilde{R}_{p-1}(z),\\
\tilde{Q}_p(z) &= (2p+1)\alpha_p q_{p-1}(z) + z\tilde{Q}_{p-1}(z),\\
q_p(z) &= 2(2p-1)q_{p-1}(z) + z^2 q_{p-2}(z),
\end{aligned}$$

together with the initial conditions $\tilde{R}_0(z) = \alpha_0$*,* $\tilde{Q}_0(z) = \alpha_0 + z$*,* $q_{-1}(z) = 1$*, and* $q_0(z) = 2 + z$*.*

Proof. Let us first consider the relation $\det(-zI - A_p - \frac{1}{\lambda}D_p) = 0$. By Theorem 1 we know that

$$\det\left(-zI - A_p - \frac{1}{\lambda}D_p\right) = Q_p(-z) - \frac{1}{\lambda}R_p(z).$$

Using the recursion relations for Q_p and R_p we obtain

$$\begin{aligned}
Q_p(-z) - \frac{1}{\lambda}R_p(z) = -zQ_{p-1}(-z) + (2p+1)\left(R_{p-1}(-z) + Q_{p-1}(-z)\right)\\
- \frac{1}{\lambda}\left(-zR_{p-1}(z) + (2p+1)(R_{p-1}(z) + Q_{p-1}(z))\right),
\end{aligned}$$

which can be rewritten,

$$\begin{aligned}
Q_p(-z) - \frac{1}{\lambda}R_p(z) = -z\left[Q_{p-1}(-z) - \frac{1}{\lambda}R_{p-1}(z)\right]\\
+ (2p+1)\Big(R_{p-1}(-z) + Q_{p-1}(-z) - \frac{1}{\lambda}\left[R_{p-1}(z) + Q_{p-1}(z)\right]\Big),
\end{aligned}$$

or, equivalently,

$$\begin{aligned}
\det\left(-zI - A_p - \frac{1}{\lambda}D_p\right) = -z\det\left(-zI - A_{p-1} - \frac{1}{\lambda}D_{p-1}\right)\\
+ (2p+1)\left(q_{p-1}(-z) - \frac{1}{\lambda}q_{p-1}(z)\right), \quad (26)
\end{aligned}$$

where $q_{p-1}(z) = Q_{p-1}(z) + R_{p-1}(z)$. Writing the matrix $\left(-zI - A_p - \frac{1}{\lambda}D_p\right)$ in full we see

$$\det\left(-zI - A_p - \frac{1}{\lambda}D_p\right)$$
$$= \begin{vmatrix}
-z + (1 - \frac{1}{\lambda}) & -(-1 + \frac{1}{\lambda}) & \dots & b_{p0} + (-1)^p - \frac{1}{\lambda}\\
3(-1 + \frac{1}{\lambda}) & -z + 3(1 + \frac{1}{\lambda}) & \dots & 3(b_{p1} + (-1)^{p+1} + \frac{1}{\lambda})\\
5(1 - \frac{1}{\lambda}) & 5(-1 - \frac{1}{\lambda}) & \dots & 5(b_{p2} + (-1)^{p+2} - \frac{1}{\lambda})\\
\vdots & \vdots & \ddots & \vdots\\
(2p+1)(-1)^p(1 - \frac{1}{\lambda}) & (2p+1)(-1)^p(-1 - \frac{1}{\lambda}) & \dots & -z + (2p+1)(1 - (-1)^p\frac{1}{\lambda})
\end{vmatrix}.$$

Using the linearity of the determinant function in the last entry we write

$$\det\left(-zI - A_p - \frac{1}{\lambda}D_p\right) = -z\det\left(-zI - A_{p-1} - \frac{1}{\lambda}D_{p-1}\right) + \det(B_p), \quad (27)$$

where

$$B_p = \begin{pmatrix} -z+(1-\frac{1}{\lambda}) & -(-1+\frac{1}{\lambda}) & \dots & b_{p0}+(-1)^p-\frac{1}{\lambda} \\ 3(-1+\frac{1}{\lambda}) & -z+3(1+\frac{1}{\lambda}) & \dots & 3(b_{p1}+(-1)^{p+1}+\frac{1}{\lambda}) \\ 5(1-\frac{1}{\lambda}) & 5(-1-\frac{1}{\lambda}) & \dots & 5(b_{p2}+(-1)^{p+2}-\frac{1}{\lambda}) \\ \vdots & \vdots & \ddots & \vdots \\ (2p+1)(-1)^p(1-\frac{1}{\lambda}) & (2p+1)(-1)^p(-1-\frac{1}{\lambda}) & \dots & (2p+1)(1-(-1)^p\frac{1}{\lambda}) \end{pmatrix}.$$

Next, by adding/subtracting an appropriate multiple of the last row of B_p to every other row we can simplify $\det(B_p)$ to

$$\det(B_p) = (2p+1)\begin{vmatrix} -z & 2 & \dots & b_{p0} \\ 0 & -z & \dots & 3b_{p1} \\ 0 & 0 & \dots & 5b_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ (-1)^p(1-\frac{1}{\lambda}) & (-1)^p(-1-\frac{1}{\lambda}) & \dots & (1-(-1)^p\frac{1}{\lambda}) \end{vmatrix}. \tag{28}$$

Comparing (27) to (26) and using (28), we obtain

$$\begin{vmatrix} -z & 2 & \dots & b_{p0} \\ 0 & -z & \dots & 3b_{p1} \\ 0 & 0 & \dots & 5b_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ (-1)^p(1-\frac{1}{\lambda}) & (-1)^p(-1-\frac{1}{\lambda}) & \dots & (1-(-1)^p\frac{1}{\lambda}) \end{vmatrix} = q_{p-1}(-z) - \frac{1}{\lambda}q_{p-1}(z). \tag{29}$$

When we return to the modified DG scheme, we write $\det\left(-z-\tilde{A}_p-\frac{1}{\lambda}\tilde{D}_p\right)$ in full to obtain,

$$\det\left(-zI-\tilde{A}_p-\frac{1}{\lambda}\tilde{D}_p\right) =$$

$$\begin{vmatrix} -z+\alpha_0(1-\frac{1}{\lambda}) & -\alpha_0(-1+\frac{1}{\lambda}) & \dots & b_{p0}+\alpha_0((-1)^p-\frac{1}{\lambda}) \\ 3\alpha_1(-1+\frac{1}{\lambda}) & -z+3\alpha_1(1+\frac{1}{\lambda}) & \dots & 3(b_{p1}+\alpha_1((-1)^{p+1}+\frac{1}{\lambda})) \\ 5\alpha_2(1-\frac{1}{\lambda}) & 5\alpha_2(-1-\frac{1}{\lambda}) & \dots & 5(b_{p2}+\alpha_2((-1)^{p+2}-\frac{1}{\lambda})) \\ \vdots & \vdots & \ddots & \vdots \\ (2p+1)(-1)^p\alpha_p(1-\frac{1}{\lambda}) & (2p+1)(-1)^p\alpha_p(-1-\frac{1}{\lambda}) & \dots & -z+(2p+1)\alpha_p(1-(-1)^p\frac{1}{\lambda}) \end{vmatrix}.$$

If we again use the linearity of the determinant function in the last entry we obtain

$$\det\left(-zI-\tilde{A}_p-\frac{1}{\lambda}\tilde{D}_p\right) = -z\det\left(-zI-\tilde{A}_{p-1}-\frac{1}{\lambda}\tilde{D}_{p-1}\right)+\det(\tilde{B}_p), \tag{30}$$

where,

$$\tilde{B}_p = \begin{pmatrix} -z+\alpha_0(1-\frac{1}{\lambda}) & -\alpha_0(-1+\frac{1}{\lambda}) & \dots & b_{p0}+\alpha_0((-1)^p-\frac{1}{\lambda}) \\ 3\alpha_1(-1+\frac{1}{\lambda}) & -z+3\alpha_1(1+\frac{1}{\lambda}) & \dots & 3(b_{p1}+\alpha_1((-1)^{p+1}+\frac{1}{\lambda})) \\ 5\alpha_2(1-\frac{1}{\lambda}) & 5\alpha_2(-1-\frac{1}{\lambda}) & \dots & 5(b_{p2}+\alpha_2((-1)^{p+2}-\frac{1}{\lambda})) \\ \vdots & \vdots & \ddots & \vdots \\ (2p+1)(-1)^p\alpha_p(1-\frac{1}{\lambda}) & (2p+1)(-1)^p\alpha_p(-1-\frac{1}{\lambda}) & \dots & (2p+1)\alpha_p(1-(-1)^p\frac{1}{\lambda}) \end{pmatrix}.$$

By again adding/subtracting an appropriate multiple of the last row of $\tilde{B}_p$ to every other row we can simplify $\det(\tilde{B}_p)$ to

$$\det(\tilde{B}_p) = (2p+1)\alpha_p\begin{vmatrix} -z & 2 & \dots & b_{p0} \\ 0 & -z & \dots & 3b_{p1} \\ 0 & 0 & \dots & 5b_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ (-1)^p(1-\frac{1}{\lambda}) & (-1)^p(-1-\frac{1}{\lambda}) & \dots & (1-(-1)^p\frac{1}{\lambda}) \end{vmatrix}, \tag{31}$$

and using (29) we obtain,

$$\det(\tilde{B}_p) = (2p+1)\alpha_p \left(q_{p-1}(-z) - \frac{1}{\lambda} q_{p-1}(z) \right). \tag{32}$$

Finally, if we define the polynomials $\tilde{Q}_p(z)$ and $\tilde{R}_p(z)$ so that

$$\det\left(-zI - \tilde{A}_p - \frac{1}{\lambda}\tilde{D}_p \right) = \tilde{Q}_p(-z) - \frac{1}{\lambda}\tilde{R}_p(z),$$

for which a direct calculation shows $\tilde{Q}_0(z) = \alpha_0 + z$ and $\tilde{R}_0(z) = \alpha_0$, we can use (32) to write (30) as

$$\tilde{Q}_p(-z) - \frac{1}{\lambda}\tilde{R}_p(z) = -z\left[\tilde{Q}_{p-1}(-z) - \frac{1}{\lambda}\tilde{R}_{p-1}(z) \right] + (2p+1)\alpha_p \left[q_{p-1}(-z) - \frac{1}{\lambda} q_{p-1}(z) \right].$$

Comparing the coefficients on λ, we arrive at the recursive relations

$$\tilde{R}_p(z) = (2p+1)\alpha_p q_{p-1}(z) - z\tilde{R}_{p-1}(z),$$
$$\tilde{Q}_p(z) = (2p+1)\alpha_p q_{p-1}(z) + z\tilde{Q}_{p-1}(z).$$

The final recursive relation, for $q_p(z)$, follows immediately from the recursion relations in Theorem 1 and the fact that $q_p(z) = Q_p(z) + R_p(z)$. □

The coefficients of the polynomials $\tilde{R}_p(z)$ and $\tilde{Q}_p(z)$ will now depend linearly on the α_m multipliers since the polynomials $q_p(z)$ are independent of α_m. From the above theorem we know that when $\alpha_m = 1, \forall m$, $\tilde{f}_p(z)$ will be the Padé approximant of e^z to order $\mathcal{O}(z^{2p+2})$. Upon choosing some coefficients $\alpha_m \neq 1$ we expect to find that

$$\tilde{f}_p(z) = e^z + \mathcal{O}(z^M)$$

where $p+1 \leq M < 2p+2$. Therefore, the original unmodified DG scheme, which results from choosing each α_m to be 1, is in a sense optimally accurate since it obtains the full $(2p+1)$-th order accuracy in dissipation and $(2p+2)$-th order accuracy in dispersion.

Let us rigorously demonstrate the effects of altering the α_m multipliers on the accuracy of the modified DG scheme through the following theorem.

Theorem 3 (mDG)**.** *Let α_m be the lowest order multiplier (smallest m) for which $\alpha_m \neq 1$ in the matrices $\tilde{A}_p$ and $\tilde{D}_p$. Then $\tilde{f}_p(z) = e^z + \mathcal{O}(z^{p+1+m})$.*

To prove this theorem, let us first establish a useful lemma.

Lemma 1. *For each p, $[q_{p-1}(z) - q_{p-1}(-z)e^z] = \mathcal{O}(z^{2p+1})$.*

Proof. From Theorems 1 and 2 we know

$$\frac{R_{p-1}(z)}{Q_{p-1}(-z)} = e^z + \mathcal{O}(z^{2p}),$$

which implies

$$R_{p-1}(z) - Q_{p-1}(-z)e^z = \mathcal{O}(z^{2p}). \tag{33}$$

Furthermore, we know

$$\frac{R_p(z)}{Q_p(-z)} = e^z + \mathcal{O}(z^{2p+2}).$$

Using the recursive relations for $\tilde{R}_p(z)$ and $\tilde{Q}_p(-z)$ in Proposition 1 with $\alpha_m = 1,\ \forall m$ we find

$$\frac{(2p+1)q_{p-1}(z) - zR_{p-1}(z)}{(2p+1)q_{p-1}(-z) - zQ_{p-1}(-z)} = e^z + \mathcal{O}(z^{2p+2}).$$

Rearranging we obtain,

$$(2p+1)(q_{p-1}(z) - q_{p-1}(-z)e^z) - z(R_{p-1}(z) - Q_{p-1}(-z)e^z) = \mathcal{O}(z^{2p+2}).$$

Using (33) we find that $z(R_{p-1}(z) - Q_{p-1}(-z)e^z)$ is $\mathcal{O}(z^{2p+1})$ and this implies that $(2p+1)(q_{p-1}(z) - q_{p-1}(-z)e^z)$ is also $\mathcal{O}(z^{2p+1})$ and has the same leading term as $z(R_{p-1}(z) - Q_{p-1}(-z)e^z)$. This establishes the result. □

We now proceed to prove the theorem.

Proof of Theorem 3. Beginning with the base case of $m = 0$ and $p = 0$, a direct calculation of $\tilde{f}_0(z)$ yields

$$\begin{aligned}\tilde{f}_0(z) &= \frac{\alpha_0}{\alpha_0 - z} \\ &= 1 + \frac{1}{\alpha_0} z + \dots \\ &= e^z + \mathcal{O}(z^1).\end{aligned}$$

Next, suppose $p > 0$ and $m \leq p$. We will use the recurrence relations for $\tilde{R}_p$ and $\tilde{Q}_p$ to calculate $\tilde{f}_p$. Since α_m is the lowest order multiplier for which $\alpha_m \neq 1$, we have that the polynomials $\tilde{R}_{m-1}$ and $\tilde{Q}_{m-1}$ are unmodified. Hence $\tilde{R}_{m-1} = R_{m-1}$, $\tilde{Q}_{m-1} = Q_{m-1}$ and $\tilde{f}(z) = f_{m-1}(z)$. From Theorem 2, $\tilde{f}_{m-1}(z) = f_{m-1}(z) = e^z + \mathcal{O}(z^{2m})$. This implies

$$\tilde{R}_{m-1}(z) - \tilde{Q}_{m-1}(-z)e^z = \mathcal{O}(z^{2m}). \tag{34}$$

We then examine the residual $\tilde{R}_m(z) - \tilde{Q}_m(-z)e^z$ and use Proposition 1 to find

$$\begin{aligned}\tilde{R}_m(z) - \tilde{Q}_m(-z)e^z &= (2m+1)\alpha_m q_{m-1}(z) - z\tilde{R}_{m-1}(z) - ((2m+1)\alpha_m q_{m-1}(-z) - z\tilde{Q}_{m-1}(-z))e^z, \\ &= (2m+1)\alpha_m(q_{m-1}(z) - q_{m-1}(-z)e^z) - z(\tilde{R}_{m-1}(z) - \tilde{Q}_{m-1}(-z)e^z), \\ &= \mathcal{O}(z^{2m+1}),\end{aligned}$$

and therefore by dividing by $\tilde{Q}_m(-z)$ we obtain that $\tilde{f}_m(z) - e^z = \mathcal{O}(z^{2m+1})$. Note that in the last line we use both (34) and Lemma 1. Also note that from Theorem 2 we know that the coefficient on z^{2m+1} in the residual of $\tilde{f}_m(z) - e^z$ must be multiplied by $(\alpha_m - 1)$ since if $\alpha_m = 1$ then $\tilde{f}_m(z) - e^z = \mathcal{O}(z^{2m+2})$. Finally, we do the inductive step by examining the residual of $\tilde{R}_{m+1}(z) - \tilde{Q}_{m+1}(-z)e^z$ to find

$$\begin{aligned}\tilde{R}_{m+1}(z) - \tilde{Q}_{m+1}(-z)e^z &= (2m+3)\alpha_{m+1} q_m(z) - z\tilde{R}_m(z) - ((2m+3)\alpha_{m+1} q_m(-z) - z\tilde{Q}_m(-z))e^z, \\ &= (2m+3)\alpha_{m+1}(q_m(z) - q_m(-z)e^z) - z(\tilde{R}_m(z) - \tilde{Q}_m(-z)e^z), \\ &= \mathcal{O}(z^{2m+3}) + \mathcal{O}(z^{2m+2}), \\ &= \mathcal{O}(z^{2m+2}).\end{aligned}$$

Hence by dividing by $\tilde{Q}_{m+1}(-z)$ we obtain $\tilde{f}_{m+1}(z) - e^z = \mathcal{O}(z^{2m+2})$. Repeating these calculations up to $\tilde{R}_p$ and $\tilde{Q}_p$ we find

$$\tilde{f}_p(z) = e^z + \mathcal{O}(z^{p+m+1}),$$

which is the desired result. □

Now that we have established the order of the error between $\tilde{f}_p(z)$ and e^z we can state our primary result on the accuracy of the modified DG scheme.

Corollary 1. *Let α_m be the lowest order multiplier (smallest m) for which $\alpha_m \neq 1$ in the matrices $\tilde{A}_p$ and $\tilde{D}_p$. Then for the non-dimensional numerical wave number, it holds*

$$K_h = K + C_1(iK)^{p+m+1} + C_2(iK)^{p+m+2} + \dots,$$

and therefore,

$$k_h = k + C_1(ik)^{p+m+1}h^{p+m} + C_2(ik)^{p+m+2}h^{p+m+1} + \dots.$$

From this corollary we can conclude that if $p+m+1$ is odd then the order of the dispersion error of the modified DG scheme (the real part of the error in k_h) is $\mathcal{O}(h^{p+m+1})$ and the order of the dissipation error (the imaginary part of the error in k_h) is $\mathcal{O}(h^{p+m})$. On the other hand, if $p+m+1$ is even then the order of the dissipation error of the modified DG scheme is $\mathcal{O}(h^{p+m+1})$ and the order of the dispersion error is $\mathcal{O}(h^{p+m})$.

Corollary 2. *If the lowest order multiplier α_0 is chosen to be not equal to one, the order of the errors in the numerical wave number k_h will be $\mathcal{O}(h^p)$. Since the order of the spatial approximation is $p+1$, the order of convergence of the scheme will be reduced to order p when $\alpha_0 \neq 1$.*

Therefore, for the remainder of this paper we will take $\alpha_0 = 1$ to preserve the usual order $p+1$ convergence rate.

Remark 1. *We remark that the proposed modifications to the numerical flux do not affect consistency of the scheme. This is because on smooth solutions the jump term is zero and will not contribute and therefore the numerical flux remains consistent with the exact flux function $f(u)$. Hence the original results established by Cockburn and Shu in [5, 4] on the $(p+1)$-th order consistency of the DG method will carry over to this modified scheme. We can therefore conclude by the equivalence theorem of Lax-Richmeyer that the modified scheme will preserve the usual $p+1$ convergence rate for linear equations, provided the scheme is linearly stable.*

For nonlinear equations the proof of the TVDM property presented in [5] can be verbatim applied to the modified scheme provided $\alpha_0 = 1$. In particular, Lemma 2.1 uses only the equation for the c_{j0} and the values of the solution at the endpoints of the interval. Since the equation c_{j0} is unmodified, and the endpoint values are limited in the same manner, the lemma holds. Moreover Lemma 2.3 in [5] will also hold with $p = 1$ and the minmod limiter. Hence the modified scheme preserves the usual order $p+1$ convergence for smooth nonlinear problems provided it is stable and $\alpha_0 = 1$.

In Proposition 2 in the next section we prove linear stability of the modified scheme in the case where only the highest order multiplier is taken not equal to one. However, when more multipliers are modified the scheme may not be linearly stable. In these cases we investigate stability by plotting the spectrum of the discrete spatial operator of the mDG scheme.

4 Stability of the mDG Method

In this section we will study what effects modifying the flux multipliers α_k will have on the linear stability of the modified DG scheme. We pair the DG spatial discretization of order p with an order $p+1$ time-integration scheme, e.g. Runge-Kutta-$(p+1)$, in order to ensure a global convergence rate of order $p+1$. For the linear advection equation, when using an explicit order $p+1$ Runge-Kutta time-integration scheme to discretize (14), it is known [5] that the stability restriction on the size of the time step Δt scales with p as

$$\Delta t \lesssim \frac{h}{a(2p+1)}.$$

This simple estimate is at most 5% smaller than the exact CFL number [6]. This time step restriction is found by choosing Δt to be small enough that the spectrum of $\frac{a\Delta t}{h}\mathbf{L}$ is contained within the absolute stability region of, in this case, Runge-Kutta-$(p+1)$. Upon altering the multipliers in the modified DG scheme (10), the spectrum of this linear operator $\frac{a\Delta t}{h}\mathbf{L}$ will be changed. It is therefore possible that this stability restriction can be relaxed by choosing the multipliers α_k in some particular way. Since determining the spectrum of this operator explicitly is very difficult, we will resort to numerically calculating its eigenvalues and determine the time step restriction by numerically searching for the largest CFL number such that the spectrum will be contained in the absolute stability region of RK-$(p+1)$. We will begin by only considering changes in the highest multiplier α_p since, as we will see, significant gains can be made in the relaxation of the stability restriction through only modifying the highest multiplier. We will then move on to study the effects of changing more than the highest multiplier.

4.1 Case 1: Only highest flux multiplier, α_p, is not equal to one.

Before we begin, let us note that in the particular case that only the highest multiplier of the modified scheme, α_p, is taken to be not equal to 1, we have a corollary of Theorem 3.

Corollary 3. *If the DG scheme is modified by only changing the highest multiplier, α_p, then the order of the dispersion error of the scheme is lowered by two to $\mathcal{O}(h^{2p})$ and the order of the dissipation error remains $\mathcal{O}(h^{2p+1})$.*

This corollary tells us that upon modifying the highest multiplier the order of accuracy in dissipation and dispersion of the scheme is only minimally affected. Therefore, the improvements in the stability restriction resulting from the modification of only the highest coefficient will have the benefit of only mildly reducing the orders of the error in dissipation and dispersion of the DG scheme. This is particularly true when using a very high order approximation since for large p the differences between an $\mathcal{O}(h^{2p+2})$ error and an $\mathcal{O}(h^{2p})$ error will be fairly negligible.

In Figure 1 we show the eigenvalues of the operator $\mathbf{L}$ for the $p = 1, 2, 3,$ and 4 schemes, respectively, with different values for the highest multiplier α_p in each case. In each figure, we show with the 'o' marker the spectrum

Figure 1: Eigenvalues of the operator $\mathbf{L}$, the spatial DG discretization for the linear advection equation, for the $p = 1$ and 2 (top) and $p = 3$, and 4 (bottom), with $N = 50$. We show in each figure the spectrum of $\mathbf{L}$ for $\alpha_p = 1, \frac{3}{2}$, and $\frac{1}{2}$.

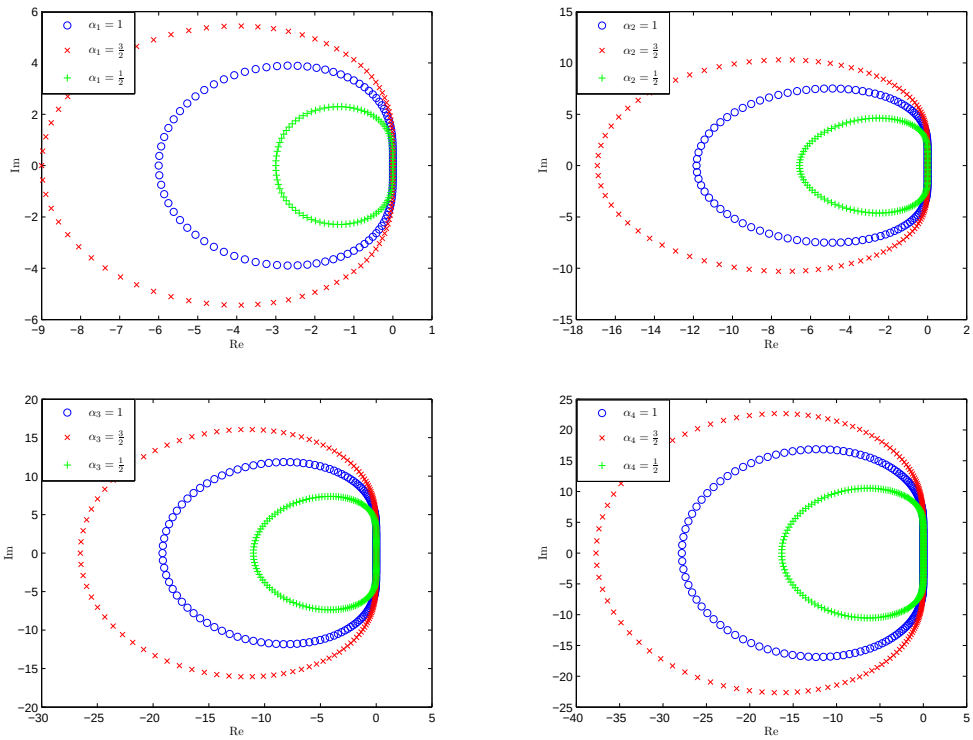


Table 1: Largest CFL numbers obtained with the modified DG scheme on the linear advection equation for $p = 1, 2, \ldots, 10$, only modifying the highest order coefficient. Relative increase is calculated as the ratio between the increased CFL of the modified scheme, divided by the CFL number of the original DG scheme.

p	α_p	CFL	Relative Increase
1	1.000	0.33	3.00
	0.333	1.00	
2	1.000	0.21	2.97
	0.210	0.62	
3	1.000	0.14	2.60
	0.260	0.37	
4	1.000	0.11	2.46
	0.270	0.28	
5	1.000	0.09	2.40
	0.330	0.22	
6	1.000	0.08	2.34
	0.345	0.19	
7	1.000	0.07	2.27
	0.360	0.16	
8	1.000	0.06	2.24
	0.380	0.14	
9	1.000	0.05	2.21
	0.385	0.12	
10	1.000	0.05	2.19
	0.395	0.11	

for $\alpha_p = 1$, which is the spectrum of the original DG scheme, together with the spectra for $\alpha_p = \frac{3}{2}$ and $\alpha_p = \frac{1}{2}$ with the 'x' and '+' markers, respectively. We notice from these figures that, in general, the modification of the highest coefficient has the effect of scaling the spectrum of the operator $\mathbf{L}$. In particular, upon increasing the α_p multiplier the spectrum of $\mathbf{L}$ is enlarged, while decreasing the α_p multiplier reduces the spectrum of $\mathbf{L}$. From this, we immediately see that when $\alpha_p < 1$, and the spectrum of $\mathbf{L}$ is reduced, we are able to choose the CFL number larger and still have a stable scheme. In contrast, when $\alpha_p > 1$ we must choose the CFL number smaller and the stability condition of the scheme is made more restrictive. Although for completeness we include the cases when $\alpha_p > 1$ in our numerical tests below, we remark that modifying the DG scheme in this way has little benefit since both the stability restriction is tightened and the accuracy of the scheme is reduced.

Now that we have established that the stability restriction of the DG scheme can be relaxed through reducing the highest multiplier α_p, our next pursuit is to determine precisely the degree to which the stability condition can be improved, what choices of α_p give us the most relaxed time-step restriction, and how much of an improvement we can expect to gain for very high order approximations. To answer these questions, we have used a MATLAB program which calculates the spectrum of $\mathbf{L}$ for varying values of α_p and uses this spectrum to find the largest CFL number so that the complete spectrum of $CFL \cdot \mathbf{L}$ is contained within the absolute stability region of RK-$(p+1)$ via a bisection algorithm. In Table 1 we present the largest CFL number we were able to obtain using this program for schemes of order $p = 1, 2, \ldots, 10$, together with the value of α_p for which the scheme obtains this CFL number. From this we see that we are able to achieve a significant increase in the usual CFL number of the DG scheme. We conjecture that for very high order schemes we can expect to obtain a two-fold increase in the CFL number of the DG scheme by only modifying the highest multiplier to be $\alpha_p \approx 0.4$. We note that this significant gain in the CFL number comes at the cost of only one order of accuracy in the form of a dispersive error, while no additional dissipative error is introduced. We can establish another property of the scheme with this modification: the semi-discrete scheme (10) is linearly stable for any choice of $\alpha_p > 0$.

Proposition 2. *The modified DG scheme* (10) *with each multiplier* $\alpha_m = 1$, $m = 1, \ldots, p-1$, *and* $\alpha_p > 0$, *is linearly stable.*

Proof. Without loss of generality, we can assume $a = 1$ in the linear advection equation. Using $\alpha_m = 1, m = 1, \ldots, p-1$, the scheme (10) with the upwind flux can be written

$$\frac{h_j}{2m+1}\dot{c}_{jm} = -(-1)^m \left[U_j(x_j) - U_{j-1}(x_j)\right] - \int_{x_j}^{x_{j+1}} \frac{dU_j}{dx} P_m \, dx, \quad m = 0, 1, \ldots, p-1, \tag{35}$$

$$\frac{h_j}{2p+1}\dot{c}_{jp} = -(-1)^p \alpha_p \left[U_j(x_j) - U_{j-1}(x_j)\right] - \int_{x_j}^{x_{j+1}} \frac{dU_j}{dx} P_p \, dx. \tag{36}$$

Multiplying each equation (35) by $c_{jm}(t)$, then multiplying (36) by $\frac{1}{\alpha_p} c_{jp}(t)$ and summing, we obtain

$$\frac{1}{2}\frac{d}{dt}\left[\left(\sum_{m=0}^{p-1} \frac{h_j}{2m+1} c_{jm}^2\right) + \frac{h_j}{(2p+1)\alpha_p} c_{jp}^2\right] = -U_j(x_j)\left[U_j(x_j) - U_{j-1}(x_j)\right] - \int_{x_j}^{x_{j+1}} \frac{dU_j}{dx}\left[\left(\sum_{m=0}^{p-1} c_{jm} P_m\right) + \frac{1}{\alpha_p} c_{jp} P_p\right] dx. \tag{37}$$

Since $\frac{dU_j}{dx}$ is a polynomial of degree less than p, the integral $\int_{x_j}^{x_{j+1}} \frac{dU_j}{dx} P_p \, dx = 0$. We then obtain

$$\begin{aligned}\int_{x_j}^{x_{j+1}} \frac{dU_j}{dx}\left[\left(\sum_{m=0}^{p-1} c_{jm} P_m\right) + \frac{1}{\alpha_p} c_{jp} P_p\right] dx &= \int_{x_j}^{x_{j+1}} \frac{dU_j}{dx}\left[\left(\sum_{m=0}^{p-1} c_{jk} P_k\right) + c_{jp} P_p\right] dx, \\ &= \int_{x_j}^{x_{j+1}} \frac{dU_j}{dx} U_j \, dx, \\ &= \frac{1}{2} U_j^2(x_{j+1}) - \frac{1}{2} U_j^2(x_j).\end{aligned} \tag{38}$$

Substituting (38) into (37) yields

$$\frac{1}{2}\frac{d}{dt}\left[\left(\sum_{m=0}^{p-1}\frac{h_j}{2m+1}c_{jm}^2\right)+\frac{h_j}{(2p+1)\alpha_p}c_{jp}^2\right] = -U_j(x_j)\left[U_j(x_j)-U_{j-1}(x_j)\right]-\frac{1}{2}U_j^2(x_{j+1})+\frac{1}{2}U_j^2(x_j),$$
$$= -\frac{1}{2}U_j^2(x_{j+1})+U_j(x_j)U_{j-1}(x_j)-\frac{1}{2}U_j^2(x_j). \tag{39}$$

Finally, summing over the entire mesh and using the periodicity of the boundary conditions yields

$$\frac{1}{2}\frac{d}{dt}\sum_{j=0}^{N}\left(\sum_{m=0}^{p-1}\frac{h_j}{2m+1}c_{jm}^2+\frac{h_j}{(2p+1)\alpha_p}c_{jp}^2\right) = \sum_{j=0}^{N}\left(-\frac{1}{2}U_j^2(x_{j+1})+U_j(x_j)U_{j-1}(x_j)-\frac{1}{2}U_j^2(x_j)\right),$$
$$= \sum_{j=0}^{N}\left(-\frac{1}{2}U_{j-1}^2(x_j)+U_j(x_j)U_{j-1}(x_j)-\frac{1}{2}U_j^2(x_j)\right),$$
$$= -\frac{1}{2}\sum_{j=0}^{N}\left(U_j(x_j)-U_{j-1}(x_j)\right)^2 \leq 0.$$

Therefore, we find that for any $\alpha_p > 0$, $\sum_{j=0}^{N}||\mathbf{c}_j||$ will be bounded, and hence the semi-discrete scheme is linearly stable. □

4.2 Several flux multipliers are not equal to one.

When several multipliers in the modified scheme (10) are taken to be not equal to one, we encounter several difficulties. Firstly, as we have established above, as we alter more multipliers the order of accuracy diminishes as we introduce larger dispersive and dissipative errors into the scheme. Secondly, the search for the choices of the multipliers which will yield the largest gain in the CFL number becomes computationally expensive. Thirdly, in our tests we observed that when more than one multiplier is modified, the operator $\mathbf{L}$ may have eigenvalues with positive real part. Therefore, a linear stability analysis of the type presented in Proposition 2 is not possible.

To understand why the scheme can become unstable, we consider the specific case when $p = 2$ and consider modifications to the second highest multiplier, α_1. Following the arguments of Proposition 1, Theorem 3, and Corollary 1, we explicitly calculate the relation between the numerical wave number K_h and the exact wave number K (for simplicity we set $\alpha_2 = 1$) to find

$$K_h = K + i\frac{\alpha_1-1}{120}K^4 - \frac{\alpha_1(\alpha_1-1)}{1200}K^5 + \mathcal{O}(K^6).$$

From this equation, we see that when $\alpha_1 < 1$ the coefficient in front of K^4 will be negative and imaginary. Since the numerical solution is of the form $\mathbf{c}_j(t) = e^{i(jK_h-\omega t)}\hat{\mathbf{c}}$, this error term will cause the magnitude of the solution to grow with j, rather than remain bounded. Hence, this negative imaginary error in K_h is the cause of the instability that can be observed when solving (10) numerically. We note that choosing $\alpha_2 > 1$ results in a stable scheme, however the spectrum of this scheme is larger than the spectrum of the original DG scheme. Hence, this choice is of little interest.

In general, we can use these expansions of K_h to determine what choices of α_k will produce an unstable scheme. For example, if we calculate the complete expansion of K_h for $p = 3$ we find

$$K_h = K + \frac{\alpha_1-1}{1680\alpha_3}K^5 + i\frac{7\alpha_3(\alpha_2-1)+3\alpha_2(\alpha_1-1)}{70560\alpha_3^2}K^6$$
$$- \frac{49\alpha_3^2(\alpha_3-1)+35\alpha_3\alpha_2(\alpha_2-1)+(147\alpha_3^2-21\alpha_3\alpha_1+15\alpha_2^2)(\alpha_1-1)}{4939200\alpha_3^3}K^7 + \mathcal{O}(K^8),$$

and the condition that the coefficient on K^6 is positive can be written

Table 2: Largest CFL numbers obtained with the modified DG scheme on the linear advection equation for $p = 3, 4$, and 5 modifying the three highest order coefficients. Relative increase is calculated as the ratio between the increased CFL of the modified scheme, divided by the CFL number of the original DG scheme.

p	α_p	α_{p-1}	α_{p-2}	CFL	Relative Increase
3	1.00	1.00	1.00	0.14	5.40
	0.04	0.39	1.15	0.78	
4	1.00	1.00	1.00	0.11	4.06
	0.04	0.41	1.16	0.47	
5	1.00	1.00	1.00	0.09	3.88
	0.07	0.52	1.16	0.36	

$$\alpha_1 \geq \frac{7\alpha_3(1-\alpha_2)}{3\alpha_2} + 1.$$

Therefore, if we alter the highest three multipliers for the $p = 3$ scheme we can expect that the scheme will be stable if this condition is met. In general, the condition that the coefficient of K^{2p} is positive in the expansion of K_h can be written

$$\alpha_{p-2} \geq \frac{(2p+1)\alpha_p(1-\alpha_{p-1})}{(2p-3)\alpha_{p-1}} + 1.$$

Hence, this condition tells us that we can expect to obtain a stable scheme when reducing the second highest multiplier, α_{p-1}, so long as the third highest multiplier, α_{p-2}, is chosen to be sufficiently large. Using this information, we again use our MATLAB program to search for the optimal choices of the three highest multipliers. More specifically, we construct a mesh of test values for α_p, α_{p-1}, and α_{p-2} and search for the specific point in this mesh which yields the largest CFL number in the modified scheme. The mesh is then refined and the process is repeated until a desired amount of accuracy for this optimal point is obtained. The obvious downside of this modification is that we must now alter the highest *three* multipliers, rather than just the highest two. This modification will therefore have a more severe effect on the overall accuracy of the scheme. We show the results of this search in Table 2 where we see that we can again substantially improve the usual CFL number of the DG scheme. However, this large increase in the CFL number appears to diminish as the order of the scheme rises, and the effects of this modification become less disruptive.

5 Numerical Examples

In this section we will apply the modified DG scheme to several test examples to confirm its convergence rate and observe the general performance of the scheme in comparison with the standard DG scheme. We will begin by testing the modified scheme with several choices of the multipliers $\alpha_m, m = 1, \ldots, p$, to show that we retain the usual $p+1$ convergence rate on smooth solutions. We will then show how the modified scheme performs for a linear problem with several different waveforms. We will also present an example where the accuracy of the modified scheme on a fine mesh is compared to the accuracy of the DG scheme on a coarse mesh, but the computational effort of both schemes is relatively equivalent. These examples are specifically chosen with initial conditions with fine structure where mesh refinement may be more beneficial to accuracy than the higher order dissipation and dispersion errors of the DG scheme. We will conclude the section by applying the modified scheme to some non-linear problems, in which we will again confirm the convergence rate and show that the demonstrated gains in the CFL condition do indeed carry over to non-linear problems.

5.1 Linear Advection Equation

Our convergence studies were done on the following initial value problem,

$$\begin{aligned} u_t + u_x &= 0, \qquad -1 \leq x < 1, \qquad t > 0, \\ u(x,0) &= u_0(x), \\ u(-1,t) &= u(1,t), \end{aligned} \tag{40}$$

Table 3: Linear advection, (40), (41). $\mathcal{L}^1$ errors ϵ_1 and convergence rates, r, for the sine wave initial condition, $p = 1$. Errors are calculated at $t = 2$, after one full period.

	$\alpha_1 = 1, CFL = \frac{1}{3}$		$\alpha_1 = \frac{4}{3}, CFL = \frac{1}{4}$		$\alpha_1 = \frac{2}{3}, CFL = \frac{1}{2}$		$\alpha_1 = \frac{1}{3}, CFL = 0.9$	
N	ϵ_1	r	ϵ_1	r	ϵ_1	r	ϵ_1	r
16	1.26e-02	-	1.97e-02	-	6.63e-03	-	2.14e-02	-
32	3.00e-03	2.07	4.88e-03	2.01	1.73e-03	1.93	5.77e-03	1.89
64	7.29e-04	2.04	1.21e-03	2.01	4.45e-04	1.96	1.47e-03	1.98
128	1.80e-04	2.02	3.02e-04	2.01	1.12e-04	1.99	3.73e-04	1.98
256	4.47e-05	2.01	7.54e-05	2.00	2.80e-05	2.00	9.39e-05	1.99

Table 4: Linear advection, (40), (41). $\mathcal{L}^1$ errors ϵ_1 and convergence rates, r, for the sine wave initial condition, $p = 2$. Errors are calculated at $t = 2$, after one full period.

	$\alpha_2 = 1, CFL = \frac{1}{5}$		$\alpha_2 = \frac{7}{5}, CFL = \frac{1}{10}$		$\alpha_2 = \frac{2}{5}, CFL = \frac{2}{5}$		$\alpha_2 = \frac{1}{5}, CFL = \frac{3}{5}$	
N	ϵ_1	r	ϵ_1	r	ϵ_1	r	ϵ_1	r
16	1.66e-04	-	1.07e-04	-	8.10e-04	-	2.44e-03	-
32	2.06e-05	3.01	1.31e-05	3.04	9.93e-05	3.03	3.02e-04	3.02
64	2.57e-06	3.00	1.62e-06	3.02	1.23e-05	3.01	3.76e-05	3.01
128	3.21e-07	3.00	2.01e-07	3.01	1.53e-06	3.01	4.70e-06	3.00
256	4.01e-08	3.00	2.51e-08	3.00	1.91e-07	3.00	5.87e-07	3.00

with

$$u_0(x) = \frac{1}{2} \sin \pi x. \tag{41}$$

In Tables 3-5 we show the results of the convergence tests for the $p = 1, 2$, and 3 schemes. In each table, we present errors ϵ_1 in the $\mathcal{L}^1$ norm at $t = 2$ after one full period on uniform meshes having 16, 32, 64, 128, and 256 elements. To obtain a proper comparison of the accuracy of the numerical solution for each choice of the α_m multipliers, the CFL number was chosen to be as large as possible, with the exception of the case $p = 1$ and $\alpha_1 = \frac{1}{3}$. In this case, a simple calculation can show that when the time step is chosen to be precisely $\Delta t = \frac{h}{a}$ this scheme will perfectly advect, i.e. with no numerical error committed, the piecewise linear numerical solution of the linear advection equation[1]. For this reason, we choose a CFL number that is slightly less than the maximum possible. In these convergence tests, when choosing the multipliers in the modified scheme to be not equal to 1, we obtain that the scheme is less accurate in terms of the $\mathcal{L}^1$ error. This is expected, since these modifications result in increased dispersion and dissipation errors as compared to the original DG scheme and these errors lead to a faster growth of the accumulated error. The temporal component of the error also increases due to a larger time step. We also see from these tables that for any stable scheme of order $p + 1$, we retain the full $p + 1$ order convergence rate regardless of the choices for the multipliers α_m, $m = 1, \ldots, p$.

Our numerical experiments revealed that when the lowest multiplier α_0 was changed, the order of convergence of the scheme was reduced by one. This was to be expected, as shown in Corollary 2, and hence was not reported.

We attempt to compare the performance of the modified scheme with the classical DG scheme in Figures 3-5.1. We use two metrics to measure effort. In the left plots we use computational complexity, which we estimate as the number of cells times the number time steps. Since the number of time steps is proportional to the number of cells divided by the CFL number, we estimate the computational complexity using N^2/CFL. In the right plots we use the CPU clock time averaged over 20 runs on a Intel i7-2600K. In each of these figures we use the same choices of multipliers in the mDG scheme as in the convergence tests in Tables 3-5. Figures 3-5.1 indicate that the modified scheme with increased CFL number has a comparable performance in terms of work N^2/CFL, only slightly better for $p = 3$ and slightly worse for $p = 2$, but performs better than the classical DG scheme in terms of run time. This seems to indicate that for very smooth linear problems there is performance benefit to the increase in CFL number.

[1] It is worth noting that for $p = 2$ we are able to construct a scheme which also perfectly advects the piecewise quadratic solution to (40) by choosing $\alpha_0 = 1$, $\alpha_1 = \frac{1}{2}$, $\alpha_2 = \frac{1}{10}$ and $CFL = 1$. However, as discussed in section 4.2, because $\alpha_1 < 1$ and $\alpha_0 = 1$, the scheme is linearly unstable for $CFL \neq 1$.

Table 5: Linear advection, (40), (41). $\mathcal{L}^1$ errors ϵ_1 and convergence rates, r, for the sine wave initial condition, $p = 3$. Errors are calculated at $t = 2$, after one full period.

	$\alpha_3 = 1, CFL = 0.14$		$\alpha_3 = 0.33, CFL = 0.35$		$\alpha_3 = 0.04, \alpha_2 = 0.39,$ $\alpha_1 = 1.15, CFL = 0.78$	
N	ϵ_1	r	ϵ_1	r	ϵ_1	r
16	3.38e-06	-	1.74e-05	-	5.15e-04	-
32	2.11e-07	4.00	1.08e-06	4.01	3.27e-05	3.97
64	1.32e-08	4.00	6.72e-08	4.00	2.04e-06	4.00
128	8.27e-10	4.00	4.20e-09	4.00	1.28e-07	4.00
256	5.17e-11	4.00	2.62e-10	4.00	7.99e-09	4.00

Figure 2: Performance comparison between DG and mDG, $p = 3$. Left figure shows accuracy compared to esitmate of work, N^2/CFL. Right figure shows accuracy compared to CPU run time.

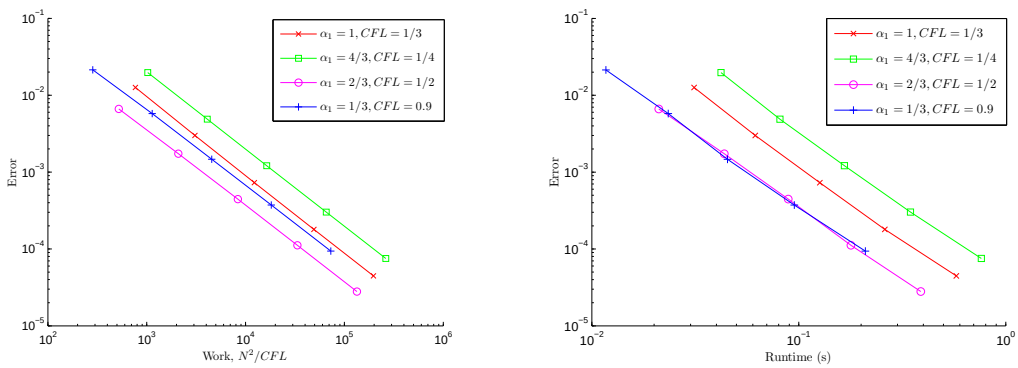


Figure 3: Performance comparison between DG and mDG, $p = 1$. Left figure shows accuracy compared to esitmate of work, N^2/CFL. Right figure shows accuracy compared to CPU run time.

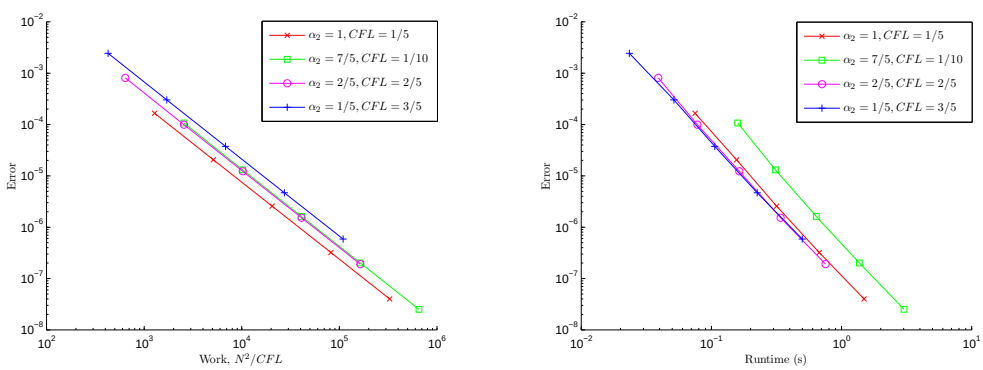


Figure 4: Performance comparison between DG and mDG, $p = 2$. Left figure shows accuracy compared to esitmate of work, N^2/CFL. Right figure shows accuracy compared to CPU run time.

5.2 Linear Advection Equation, Discontinuous Solutions

The next test - with which we can more directly observe the effects of modifying the DG scheme on a variety of waveforms - involves solving (40) with the following initial conditions [9]:

$$u_0(x) = \begin{cases} \frac{1}{6}(G(x,\beta,z-\delta) + G(x,\beta,z+\delta) + 4G(x,\beta,z)) & -0.8 \le x \le -0.6, \\ 1 & -0.4 \le x \le -0.2, \\ 1 - |10(x-0.1)| & 0 \le x \le 0.2, \\ \frac{1}{6}(F(x,\alpha,a-\delta) + F(x,\alpha,a+\delta) + 4F(x,\alpha,z)) & 0.4 \le x \le 0.6, \\ 0 & \text{otherwise}, \end{cases} \tag{42a}$$

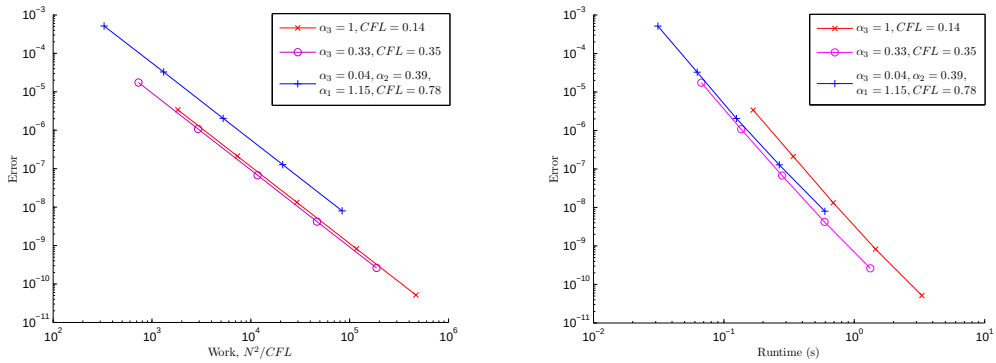
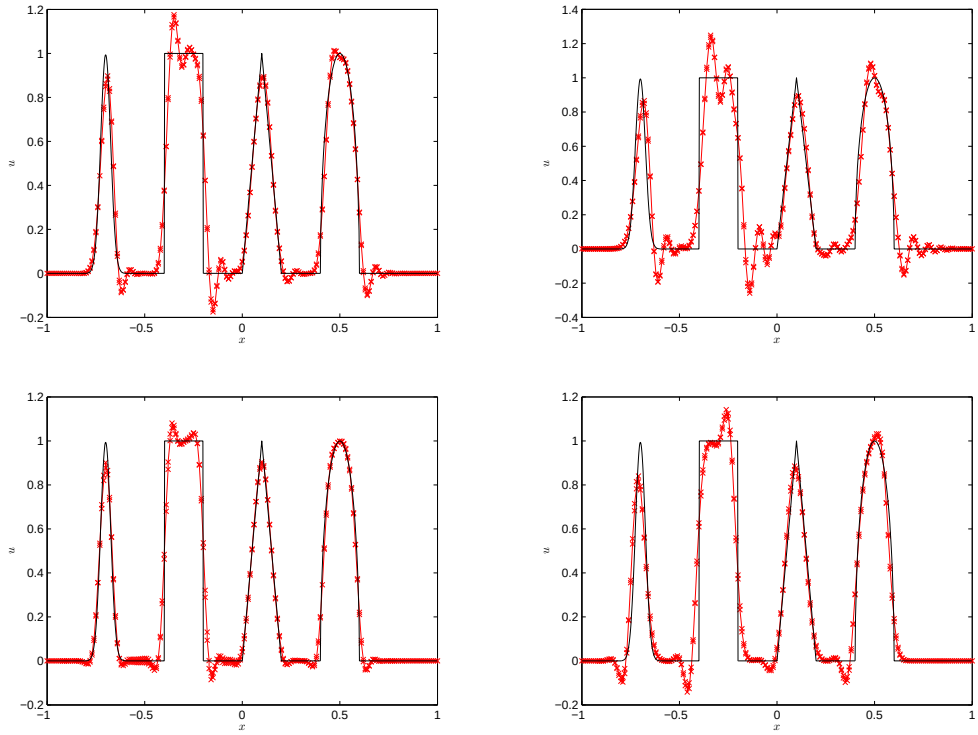


Figure 5: Linear advection, (40), (42), $p = 1$ on a mesh of $N = 200$ elements. Shown at $t = 2$, after one full period. Solid line shows the exact solution, line with 'x' markers shows the numerical solution. Top left: $\alpha_1 = 1, CFL = \frac{1}{3}$, Top Right: $\alpha_1 = \frac{4}{3}, CFL = \frac{1}{4}$, Bottom left: $\alpha_1 = \frac{2}{3}, CFL = \frac{1}{2}$, Bottom right: $\alpha_1 = \frac{1}{3}, CFL = 0.9$.

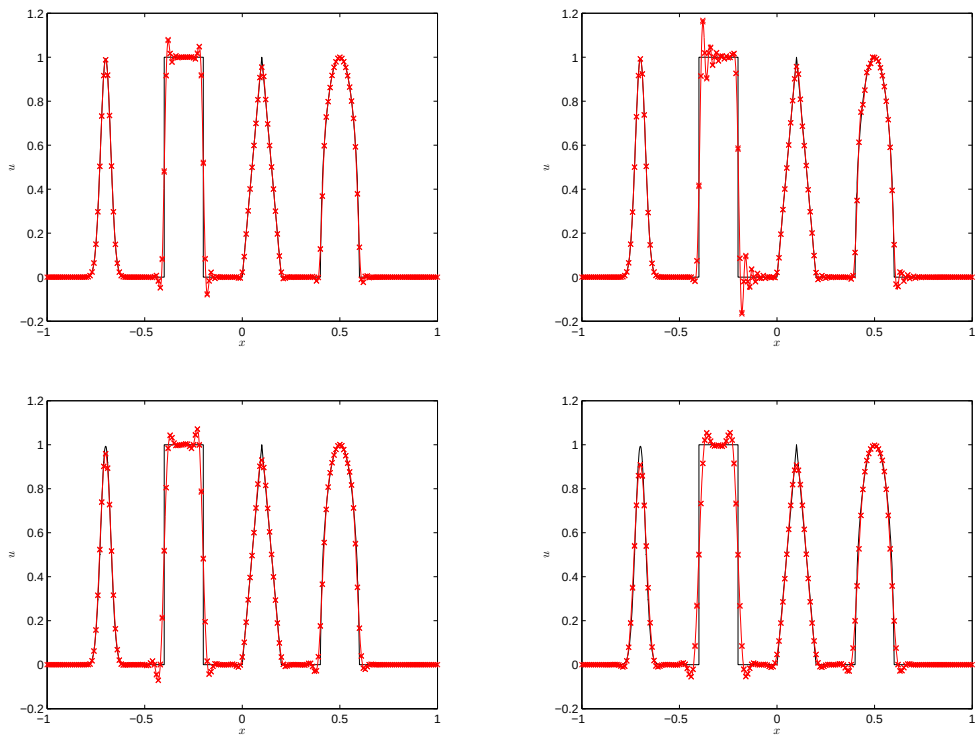


$$G(x, \beta, z) = e^{-\beta(x-z)^2}, \tag{42b}$$

$$F(x, \alpha, a) = \sqrt{\max(1 - \alpha^2(x-a)^2, 0)}, \tag{42c}$$

where $a = 0.5$, $z = -0.7$, $\delta = 0.005$, $\alpha = 10$, and $\beta = \frac{\log 2}{36\delta^2}$. This initial profile consists of a combination of Gaussians, a square pulse, a sharp triangle, and a combination of half-ellipses. We present the results with out limiting in order to discuss the effect of the induced dispersive and dissipative errors in the modified scheme. These effects are better seen in the spurious oscillations near solution discontinuities - which limiting would destroy - and in the dissipation of local extrema, to which limiters heavily contribute. We then present an example where the limiter has been applied and note that there is little difference between the schemes in terms of accuracy. Implementation of limiters, e.g. the minmod [5] or moment limiter [10], is straightforward and analogous to their implementation in classical DG schemes.

Figure 6: Linear advection, (40), (42), $p = 2$ on a mesh of $N = 200$ elements. Shown at $t = 2$, after one full period. Solid line shows the exact solution, line with 'x' markers shows the numerical solution. Top left: $\alpha_2 = 1, CFL = \frac{1}{5}$, Top Right: $\alpha_2 = \frac{7}{5}, CFL = \frac{1}{10}$, Bottom left: $\alpha_2 = \frac{2}{5}, CFL = \frac{2}{5}$, Bottom right: $\alpha_2 = \frac{1}{5}, CFL = \frac{3}{5}$.



The results of test (40)-(42) for the $p = 1, 2$, and 3 schemes are shown in Figures 5-7 at $t = 2$ after one full period, on a uniform mesh of $N = 200$ cells. In each figure we show several choices of the highest multiplier α_p and for the $p = 3$ scheme in Figure 7 we show an example where the three highest multipliers have been modified to their optimal values listed in Table 2. In Figure 5, we observe a slight shift to the left and right for $\alpha_1 = \frac{1}{3}$ and $\alpha_1 = \frac{4}{3}$, respectively, of the entire wave front for the $p = 1$ scheme. This is especially noticeable for the Gaussians and ellipses. The modified scheme for which $\alpha_1 = \frac{2}{3}$ is visually closer to the original DG scheme. This can be explained once we explicitly calculate the expansion of the numerical wave number K_h in terms of the exact wave number K from Corollary 1 for the $p = 1$ scheme,

$$K_h = K - \frac{\alpha_1 - 1}{12\alpha_1} K^3 + \mathcal{O}(K^4). \tag{43}$$

Hence, choosing $\alpha_1 > 1$ will introduce an additional dispersive error of negative sign into the usual DG scheme. On the other hand, decreasing α_1 to $\frac{2}{3}$ introduces an additional positive dispersive error. Note that the numerical wave speed, a_h, is given by $a_h = \omega/K_h$. Hence, $K_h < K$ results in $a_h > a$ and we observe numerical waves travelling slightly faster than the exact one. Similarly for $K_h > K$ we observe numerical waves travelling slightly slower.

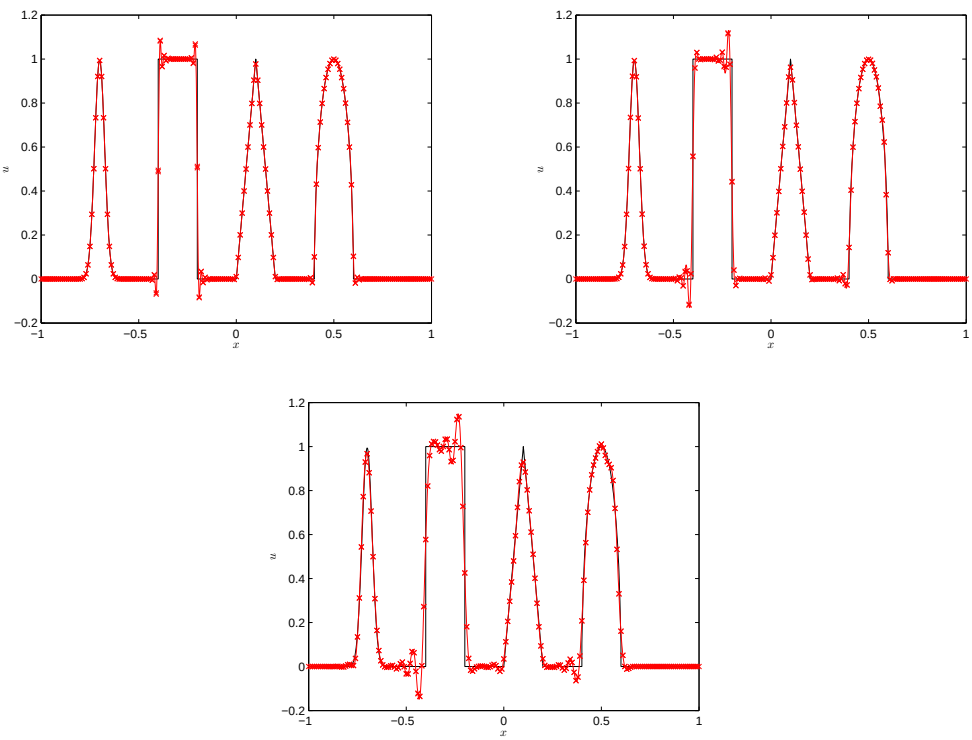
This property is true in general for the modified scheme, i.e. in the expansion of K_h for the order p scheme, when each multiplier is taken to be equal to one except the highest, the coefficient of K^{2p+2} will have a similar form to (43). Therefore, choosing $\alpha_p > 1$ will add a negative dispersive error and shift the wave fronts to the right, while choosing $\alpha_p < 1$ will add a positive dispersive error and shift the wave fronts to the left. For example, the full expansion of K_h in the $p = 2$ scheme is calculated to be

$$K_h = K + i\frac{\alpha_1 - 1}{120\alpha_2} K^4 - \frac{5\alpha_2(\alpha_2 - 1) + 3\alpha_1(\alpha_1 - 1)}{3600\alpha_2^2} K^5 + \mathcal{O}(K^6), \tag{44}$$

and therefore when $\alpha_1 = 1$,

$$K_h = K - \frac{\alpha_2 - 1}{720\alpha_2} K^5 + \mathcal{O}(K^6), \tag{45}$$

Figure 7: Linear advection, (40), (42), $p = 3$ on a mesh of $N = 200$ elements. Shown at $t = 2$, after one full period. Solid line shows the exact solution, line with 'x' markers shows the numerical solution. Top left: $\alpha_3 = 1, CFL = 0.14$, Top Right: $\alpha_3 = 0.33, CFL = 0.36$, Bottom: $\alpha_3 = 0.04, \alpha_2 = 0.39, \alpha_1 = 1.15, CFL = 0.78$.



and the effects of altering α_2 in the $p = 2$ scheme will be analogous to the effects of altering α_1 in the $p = 1$ scheme.

We note that although the order of the leading errors of K_h may stay the same for different choices of the multipliers in (43)-(45), the magnitude of the error changes with different choices. Indeed from these examples it is clear that although the formal order of accuracy remains the same, larger modifications may introduce larger errors in accuracy. In practice, care should be taken to choose the multipliers to obtain a balance between the stability gains and the deteriorating effects of the loss of accuracy.

Finally, we show in Figure 8 the results of this test for $p = 1$ with a minmod limiter implemented. We measure the errors to be 0.070, 0.079, 0.068, 0.117 for the DG, mDG with $\alpha_1 = 4/3, 2/3, 1/3$, respectively. Visually the solutions look similar, with the expection of the $\alpha_1 = 1/3$ case where the error is greater. This would seem to indicate that in the presence of discontinuities when a limiter is used there is little difference in accuracy of the solutions, i.e. for non-smooth problems the accuracy is almost completely determined by the limiter. Hence the limiter would appear to remove the detrimental effects on accuracy introduced by the modifications and the performance benefits of the modified scheme are immediate.

5.3 Burgers' Equation

To test the modified scheme on a non-linear problem, we consider Burgers' equation,

$$u_t + uu_x = 0, \tag{46}$$

on $[-1, 1]$, with periodic boundary conditions and with the sine wave initial condition, (41). We perform our convergence tests on this problem for the $p = 1$ and $p = 2$ schemes for various choices of the multipliers α_m and show the results in Tables 6 and 7. We use the same choices of multipliers as in Section 5.1 above), and present errors ϵ_1 in the $\mathcal{L}^1$ norm at $t = 0.3$, before the shock wave has formed. No limiter is used in these tests. From these tables we see that the modified scheme indeed retains the usual order of convergence for this nonlinear problem, for any

Figure 8: Linear advection, (40), (42), $p = 1$ on a mesh of $N = 200$ elements with minmod limiter. Shown at $t = 2$, after one full period. Solid line shows the exact solution, line with 'x' markers shows the numerical solution. Top left: $\alpha_1 = 1, CFL = \frac{1}{3}$, Top Right: $\alpha_1 = \frac{4}{3}, CFL = \frac{1}{4}$, Bottom left: $\alpha_1 = \frac{2}{3}, CFL = \frac{1}{2}$, Bottom right: $\alpha_1 = \frac{1}{3}, CFL = 0.9$.

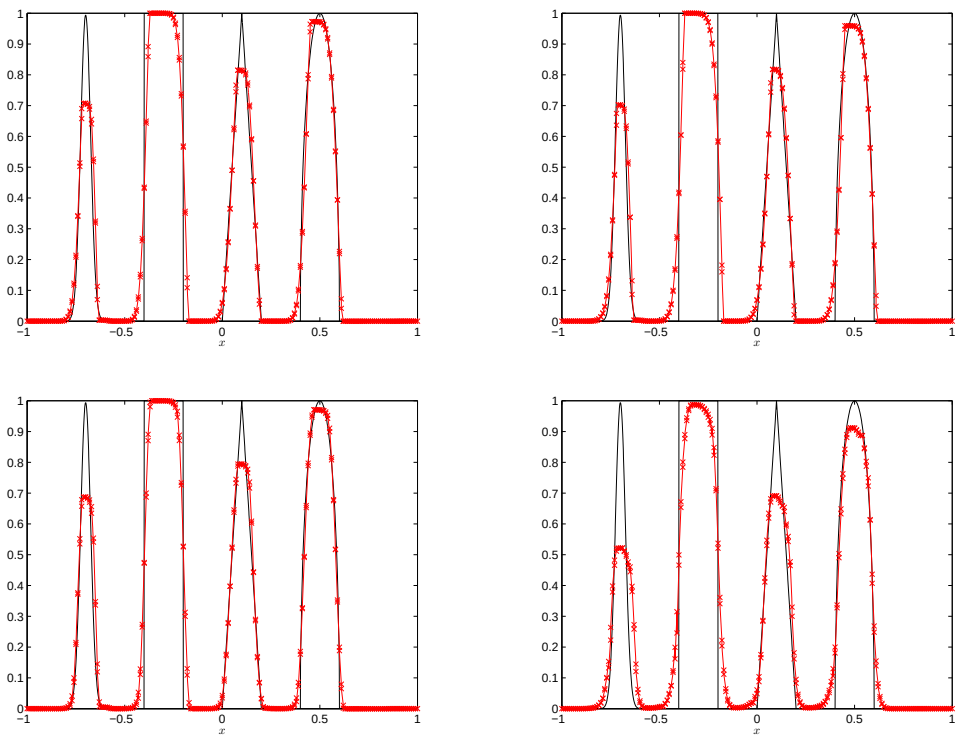


Table 6: Burgers' equation (46), (41). $\mathcal{L}^1$ errors ϵ_1 and convergence rates, r, $p = 1$. Errors are calculated at $t = 0.3$, before a shock wave forms.

	$\alpha_1 = 1, CFL = \frac{1}{3}$		$\alpha_1 = \frac{4}{3}, CFL = \frac{1}{4}$		$\alpha_1 = \frac{2}{3}, CFL = \frac{1}{2}$		$\alpha_1 = \frac{1}{3}, CFL = 0.9$	
N	ϵ_1	r	ϵ_1	r	ϵ_1	r	ϵ_1	r
16	3.83e-03	-	3.54e-03	-	5.79e-03	-	1.58e-02	-
32	1.17e-03	1.71	9.92e-04	1.83	1.74e-03	1.73	3.78e-03	1.58
64	3.24e-04	1.84	2.67e-04	1.89	4.99e-04	1.81	1.10e-03	1.78
128	8.63e-05	1.91	7.01e-05	1.93	1.37e-04	1.87	3.02e-04	1.87
256	2.24e-05	1.95	1.80e-05	1.96	3.58e-05	1.93	7.96e-05	1.93

choices of the multipliers α_m. We again observe that the performance of the DG scheme is roughly the same with that of the mDG method with increased CFL number for a fixed computation effort.

5.4 Euler Equations

To test the modified DG method for a system of equations, we consider the Euler equations, $\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = 0$ with

$$\mathbf{u} = (\rho, \rho q, E)^T, \quad \mathbf{f}(\mathbf{u}) = q\mathbf{u} + (0, P, qP)^T, \tag{47a}$$

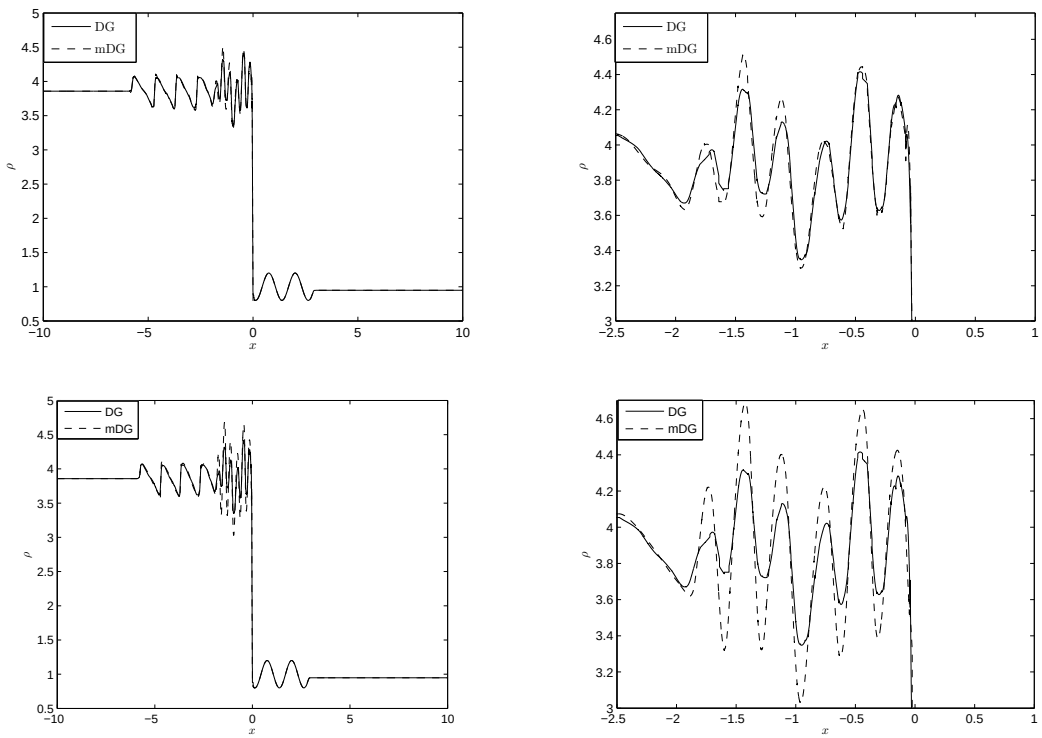
and an equation of state

$$P = (\gamma - 1)\left(E - \frac{1}{2}\rho q^2\right), \tag{47b}$$

Table 7: Burgers' equation (46), (41). $\mathcal{L}^1$ errors ϵ_1 and convergence rates, r, $p = 2$. Errors are calculated at $t = 0.3$, before a shock wave forms.

	$\alpha_2 = 1, CFL = \frac{1}{5}$		$\alpha_2 = \frac{7}{5}, CFL = \frac{1}{10}$		$\alpha_2 = \frac{2}{5}, CFL = \frac{2}{5}$		$\alpha_2 = \frac{1}{5}, CFL = \frac{3}{5}$	
N	ϵ_1	r	ϵ_1	r	ϵ_1	r	ϵ_1	r
16	2.58e-04	-	2.02e-04	-	7.04e-04	-	1.40e-03	-
32	3.43e-05	2.91	2.76e-05	2.87	9.45e-05	2.90	1.95e-04	2.84
64	4.63e-06	2.89	3.56e-06	2.95	1.22e-05	2.95	2.62e-05	2.90
128	6.16e-07	2.91	4.54e-07	2.97	1.60e-06	2.93	3.49e-06	2.91
256	8.03e-08	2.94	5.78e-08	2.97	2.10e-07	2.93	4.61e-07	2.92

Figure 9: Euler equations, (47)- (48), $p = 2$, shown at $t = 2$. Top: DGM and mDGM with $\alpha_2 = \frac{1}{5}$ on a mesh of $N = 500$ elements. Bottom: DGM on a 500-element mesh and mDGM with $\alpha_2 = \frac{1}{5}$, on a mesh of $N = 866$ elements. Right plots are zooms of left plots.



for which we take $\gamma = 1.4$, and subject to the initial data [16]

$$(\rho, q, P)(x, 0) = \begin{cases} (3.857143, -0.920279, 10.333333), & x \leq 0, \\ (1 + 0.2\sin(5x), -3.549648, 1.000000), & 0 < x < 10, \\ (1.000000, -3.549648, 1.000000), & x \geq 10. \end{cases} \tag{48}$$

This example involves the interaction of a stationary shock at $x = 0$ with a leftward-moving flow having a sinusoidal density variation. As the density perturbation passes through the shock, it produces oscillations developing into shocks of smaller amplitude. We choose this test problem since it gives us a good example of the interaction between a shock and the fine structure of the produced oscillations. In our tests we chose to use the moment limiter [10]. In Figure 9, we present the numerical solutions of the $p = 2$ scheme at $t = 2$. In the top left figure we show the unmodified DG scheme, $\alpha_2 = 1$ with $CFL = \frac{1}{5}$, and the modified scheme with $\alpha_2 = \frac{1}{5}$ and $CFL = \frac{3}{5}$, on a mesh of $N = 500$ elements. In the top right figure we show a zoomed view of the fine structure of the solution to the

left of the shock wave. In each figure we show the schemes together with a reference solution computed using the DG scheme with $p = 2$ and $N = 2500$ with the moment limiter. Surprisingly, the mDG solution is more accurate, i.e. suffers from less numerical diffusion. While a rigorous explanation of this is still an open question, one possible explanation is that the limiter destroys some of the accuracy of the fine structure at each iteration. Hence, since the modified solution is obtained using a larger time-step, the solution is less damaged by the limiter and is able to better resolve the fine structure to the left of the shock wave. Finally, in the bottom left figure we show again the unmodified DG scheme, $\alpha_2 = 1$ with $CFL = \frac{1}{5}$, on the same mesh of $N_{DG} = 500$ elements, together with the modified scheme with $\alpha_2 = \frac{1}{5}$ and $CFL = \frac{3}{5}$ on a mesh of $N_{mDG} = 866 \approx \sqrt{3}N_{DG}$ elements. The bottom right figure shows a zoomed view of the fine structure of the solution. This example demonstrates the increase in accuracy we can obtain by implementing the modified DG scheme on a refined mesh, for equivalent computation effort.

6 Conclusions

In this paper, we have proposed a family of numerical schemes obtained through a modification of the discontinuous Galerkin finite element method. It is known that the choice of numerical flux influences the spectrum of the DG scheme. For example, the central flux results in the spectrum being entirely located on the imaginary axis and the upwind flux produces a spectrum which lies in the left half-plane and grows with order of approximation p. Here, we propose a modification to the DG scheme which does not change the type of flux, but rather alters the contribution of this flux to the solution coefficients c_{jm}. This modification is obtained by multiplying the jump contributions of the numerical flux for the solution coefficient c_{jm} by a multiplier α_m. Since for one-dimensional problems, with a basis of Legendre polynomials, the coefficient c_{jm} is a numerical approximation of the m-th derivative of the solution on cell j, scaled by $C_m h^m$ where C_m is a constant, our method modifies the amount of numerical flux that is being contributed the m-th derivative of the solution. In the specific case that $\alpha_m = 1, \forall m$, we obtain the usual DG method.

The results of our study of this modified method can be summarized as follows. Firstly, the modification of the lowest order coefficient α_0 in the order p scheme immediately results in a severe accuracy loss and the order of convergence of the scheme is reduced by one. We therefore avoid such a modification and focus on modifying only the equations for the higher order coefficients of the scheme. Secondly, by analyzing how the modified scheme performs on the linear advection equation we can establish that when the coefficient α_m is modified the order of accuracy of dispersion and dissipation of the scheme is $p + m$, i.e. the accuracy is reduced from the usual accuracy of order $2p + 1$ in dissipation and order $2p + 2$ in dispersion. This reduction of accuracy introduces additional dispersive and diffusive errors to the numerical solution. Thirdly, when modifying only the highest multiplier we can prove that the method is linearly stable for any choice of α_p. Furthermore, we can expect to obtain a more relaxed stability restriction by choosing $\alpha_p \approx 0.4$. The relaxed condition allows us to take a time step twice as large, compared to the usual DG method. Finally, more multipliers may be altered and a larger improvement in the usual CFL number can be made for specific choices of α_m, but as more multipliers are altered the accuracy of the scheme is reduced as more dispersion and dissipation errors are added. Additionally, the increased time step introduces a larger temporal error into the solution.

We present a number of numerical experiments demonstrating the performance of the mDG method. In our examples, the mDG method preserves the convergence rate of the original DG method in the usual L^1 norm. For the linear advection equation with a very smooth profile, the mDG method performs similarly to the DG method for a fixed computational effort, i.e. the number of cells times the number of time steps, but outperforms the DGM in terms of CPU runtime. On the other hand, when the solution has discontinuities and limiters are applied, the mDG scheme provides a comparable solution on the same mesh, but in less computation time. Additionally fewer time steps results in less limiting which can result in fine structures of the solution from being overly smoothed by the limiter. In particular, for the Euler equations example, the mDG method results in a better solution with the CFL number being three times larger than in the usual DG method.

It is hoped that further study will illuminate a better understanding of the effects of these modifications to the DG method. In particular, more testing is necessary to determine what choices of the multipliers will be optimal in the sense of the trade-off between accuracy and the CFL number. It would also be useful to compare the mDG scheme with finite-volume and finite-difference schemes in terms of accuracy. Additionally, the results in Tables 1 and 2 indicate that there may be a pattern in the choices of the multipliers α_m which give us the largest CFL improvement, as p increases. This suggests that these choices may be related to some specific rational approximation of $\exp(z)$.

Further, the optimal choices of the multipliers α_m should also be investigated with the application of different limiters in the presence of shock waves. Finally, the extension of this approach to higher dimensional problems is straightforward in terms of implementation, however the analysis becomes complex. Preliminary numerical tests indicate that the CFL number can relaxed on two-dimensional unstructured triangular grids in an analogous way and this is the subject of current research.

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] B. Cockburn, S.Y. Lin, and C.-W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin methods for scalar conservation laws III: One dimensional systems. *Journal of Computational Physics*, 84:90–113, 1989.

[5] B. Cockburn and C.-W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin methods for scalar conservation laws II: General framework. *Mathematics of Computation*, 52:411–435, 1989.

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

[7] 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.

[8] F. Q. Hu, M.Y. Hussaini, and P. Rasetarinera. An analysis of the discontinuous galerkin method for wave propagation problems. *Journal of Computational Physics*, 151(2):921 – 946, 1999.

[9] G.-S. Jiang and C.-W. Shu. Efficient implementation of weighted ENO schemes. *Journal of Computational Physics*, 126:202–228, 1996.

[10] L. Krivodonova. Moment limiters for discontinuous Galerkin methods. *Journal of Computational Physics*, 226:276–296, 2007.

[11] L. Krivodonova and R. Qin. An analysis of the spectrum of the discontinuous Galerkin method. *Applied Numerical Mathematics*, 64:1–18, 2013.

[12] H. Luo, J. Baum, and R. Lohner. On the computation of steady-state compressible flows using a discontinuous Galerkin method. *International Journal for Numerical Methods in Engineering*, 73(5):597–623, 2008.

[13] J. Niegemann, R. Diehl, and K. Busch. Efficient low-storage rungekutta schemes with optimized stability regions. *Journal of Computational Physics*, 231(2):364 – 372, 2012.

[14] M. Parsani, D. I. Ketcheson, and W. Deconinck. Optimized explicit runge–kutta schemes for the spectral difference method applied to wave propagation problems. *SIAM Journal on Scientific Computing*, 35(2):A957–A986, 2013.

[15] S. Sherwin. Dispersion analysis of the continuous and discontinuous Galerkin formulations. In *Discontinuous Galerkin methods (Newport, RI, 1999)*, volume 11 of *Lect. Notes Comput. Sci. Eng.*, pages 425–431. Springer, Berlin, 2000.

[16] C.-W. Shu and S. Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes, II. *Journal of Computational Physics*, 83:32–78, 1989.

[17] T. Toulorge and W. Desmet. Optimal rungekutta schemes for discontinuous galerkin space discretizations applied to wave propagation problems. *Journal of Computational Physics*, 231(4):2067 – 2091, 2012.

[18] T. Warburton and T. Hagstrom. Taming the CFL number for discontinuous Galerkin methods on structured meshes. *SIAM J. Numer. Anal.*, 46(6):3151–3180, 2008.