Maximum principle preserving time implicit DGSEM for linear scalar hyperbolic conservation laws

Riccardo Milani, Florent Renac, Jean Ruel

Maximum principle preserving time implicit DGSEM for linear scalar hyperbolic conservation laws

Riccardo Milani[a], Florent Renac[a,*], Jean Ruel[a]

[a]*DAAA, ONERA, Université Paris Saclay, F-92322 Châtillon, France*

Abstract

The properties of the high-order discontinuous Galerkin spectral element method (DGSEM) with implicit backward Euler time stepping are investigated for the approximation of hyperbolic linear scalar conservation equation in multiple space dimensions. We first prove that the DGSEM scheme in one space dimension preserves a maximum principle for the cell-averaged solution when the time step is large enough. This property however no longer holds in multiple space dimensions and we propose to use the flux-corrected transport (FCT) limiting [5] based on a low-order approximation using graph viscosity to impose a maximum principle on the cell-averaged solution. These results allow us to use a linear scaling limiter [58] in order to impose a maximum principle at nodal values within elements, while limiting the cell average with the FCT limiter improves the accuracy of the limited solution. Then, we investigate the inversion of the linear systems resulting from the time implicit discretization at each time step. We prove that the diagonal blocks are invertible and provide efficient algorithms for their inversion. Numerical experiments in one and two space dimensions are presented to illustrate the conclusions of the present analyses.

Keywords: hyperbolic scalar equations, maximum principle, discontinuous Galerkin method, summation-by-parts, backward Euler

1. Introduction

We aim at developing an accurate and robust approximation of the following problem with a linear scalar advection equation with constant coefficients in $d \geq 1$ space dimensions:

$$\partial_t u + \nabla \cdot (\mathbf{c}u) = 0, \quad \text{in } \Omega \times (0, \infty), \tag{1a}$$

$$u(\cdot, 0) = u_0(\cdot), \quad \text{in } \Omega, \tag{1b}$$

with $\Omega \subset \mathbb{R}^d$, appropriate boundary conditions on $\partial\Omega$, and u_0 in $L^\infty(\mathbb{R}^d, \mathbb{R})$. Without loss of generality, we assume $\mathbf{c}$ in $\mathbb{R}^d_+$, a negative component being handled by reverting the corresponding space direction.

Problem (1) has to be understood in the sense of distributions where we look for weak solutions. Introducing the square entropy $\eta(u) = \frac{u^2}{2}$ and associated entropy flux $\mathbf{q}(u) = \mathbf{c}\frac{u^2}{2}$ pair, solutions to (1a) also satisfy

$$\partial_t \eta(u) + \nabla \cdot \mathbf{q}(u) \leq 0, \quad \text{in } \Omega \times (0, \infty), \tag{2}$$

in the sense of distributions. For compactly supported solutions, this brings uniqueness and L^2 stability. Solutions to (1) also satisfy a maximum principle:

$$m \leq u_0(x) \leq M \text{ in } \Omega \quad \Rightarrow \quad m \leq u(x,t) \leq M \text{ in } \Omega \times (0, \infty), \tag{3}$$

almost everywhere, which brings L^∞ stability.

*Corresponding author. Tel.: +33 1 46 73 37 44; fax.: +33 1 46 73 41 66.

Email addresses: riccardo.milani@onera.fr (Riccardo Milani), florent.renac@onera.fr (Florent Renac), jean.ruel@ens-paris-saclay.fr (Jean Ruel)

We are interested in the approximation of (1) with a high-order space discretization that satisfies the above properties at the discrete level. We consider the discontinuous Galerkin spectral element method (DGSEM) based on collocation between interpolation and quadrature points [26] and tensor products of one-dimensional (1D) function bases and quadrature rules. The collocation property of the DGSEM in addition to tensor-product evaluations drastically reduces the number of operations in the operators implementing the discretization and makes the DGSEM computationally efficient. Moreover, using diagonal norm summation-by-parts (SBP) operators and the entropy conservative numerical fluxes from Tadmor [48], semi-discrete entropy conservative finite-difference and spectral collocation schemes have been derived in [16, 7] and applied to a large variety of nonlinear conservation laws [19, 4, 13, 41, 40, 43, 54, 37, 58], nonconservative hyperbolic systems and balance laws [32, 42, 11, 2, 53, 52], among others.

Most of the time, these schemes are analyzed in semi-discrete form for which the time derivative is not discretized, or when coupled with explicit in time discretizations. Time explicit integration may however become prohibitive for long time simulations or when looking for stationary solutions due to the strong CFL restriction on the time step which gets smaller as the approximation order of the scheme increases to ensure either linear stability [17, 3, 27], or positivity of the approximate solution [58, 59]. The DGSEM also presents attractive features for implicit time stepping. First, the collocation property reduces the connectivity between degrees of freedom (DOFs) which makes the DGSEM well suited due to a reduced number of entries in the Jacobian matrix of the space residuals. This property has been used in [45] to rewrite the time implicit discretization of the compressible Navier-Stokes equations as a Schur complement problem at the cell level that is then efficiently solved using static condensation. Then, tensor-product bases and quadratures have motivated the derivation of tensor-product based approximations of the diagonal blocks of the Jacobian matrix by Kronecker products [51, 50] of 1D operators using singular value decomposition of a shuffled matrix [35], or a least squares alternatively in each space direction [14].

We here consider and analyze a DGSEM discretization in space associated with a first-order backward Euler time integration which allows to circumvent the CFL condition for linear stability and makes it well adapted for approximating stationary solutions or solutions containing low characteristic frequency scales. It is however of strong importance to also evaluate to what extent other properties of the exact solution are also satisfied at the discrete level. Preserving invariant domains of the equations at the discrete level is an essential property that may be required for stability of the computations. Little is known about the properties of time implicit DGSEM schemes, apart from the entropy stability which holds providing the semi-discrete scheme is entropy stable due to the dissipative character of the backward Euler time integration. An analysis of a time implicit discontinuous Galerkin (DG) method with Legendre basis functions for the discretization of a 1D linear scalar hyperbolic equation has been performed in [39] and showed that a lower bound on the time step is required for the cell-averaged solution to satisfy a maximum principle at the discrete level. A linear scaling limiter of the DG solution around its cell-average [58] is then used to obtain a maximum principle preserving scheme. Numerical experiments with linear and also nonlinear hyperbolic scalar equations and systems support the conclusion of this analysis. The theoretical proof of this lower bound uses the truncated expansion of the Dirac delta function in Legendre series that is then used as a test function in the DG scheme to prove that the Jacobian matrix of the cell-averaged discrete scheme is an M-matrix. It is however difficult to use this trick in the DGSEM scheme that uses lower-order quadrature rules and whose form is directly linked to the particular choice of Lagrange interpolation polynomials as test functions. Unfortunately, this discrete preservation of the maximum principle or positivity no longer holds in general in multiple space dimensions even on Cartesian grids and solutions with negative cell-average in some cell can be generated [30]. In the case of linear hyperbolic equations and radiative transfer equations, Ling et al. [30] showed that it is possible to impose positivity of the solution providing the approximation polynomial space is enriched with additional functions. The use of reduced order quadrature rules and suitable test functions were proposed in [55] to define a conservative scheme that preserves positivity in the case of stationary linear hyperbolic conservation laws, The work in [56] proposes limiters that allow to ensure positivity of stationary solutions of the radiative transfer equations, while keeping a particular local conservation property for stationary conservation laws. These modifications seem difficult to be directly applied to the DGSEM without loosing the collocation which is essential for the efficiency of the method. A limiter for time implicit DG schemes for nonlinear scalar equations has been proposed in [49] by reformulating the discrete problem as a constrained optimization problem and introducing Lagrange multipliers associated to the constraints. This however results in a nonlinear and nonsmooth algebraic system of equations that requires an adapted Newton method for its resolution.

In the present work, we propose an analysis of the DGSEM scheme with backward Euler time stepping for linear

hyperbolic equations on Cartesian grids. We first analyze the discrete preservation of the maximum principle property and show that it holds for the cell-averaged solution in one space dimension for sufficiently large time steps. This result is similar to the one obtained in [39] for a modal DG scheme with Legendre polynomials, though the conditions on the time step are different. The proof relies on the nilpotent property of the discrete derivative matrix evaluating the derivatives of the Lagrange interpolation polynomials at quadrature points. This property allows to easily invert the mass and stiffness matrices and derive a scheme for the cell-averaged scheme, thus allowing to derive conditions for the matrix of the associated linear system to be an M-matrix. The DOFs are then limited with the linear scaling limiter from [58] to impose a maximum principle to the whole solution. Unfortunately, this property no longer holds in multiple space dimensions similarly to the modal DG scheme [30]. We thus follow [20, 15] that propose to use the flux-corrected transport (FCT) limiter [5, 57] combining a low-order and maximum principle preserving scheme with the high-order DGSEM. The low-order scheme is obtained by adding graph viscosity [22, 21, 34] to the DGSEM scheme. The FCT limiter is here designed to preserve a maximum principle for the cell-averaged solution, not for all the DOFs. Here again, the linear scaling limiter is applied after the FCT limiter to ensure the maximum principle on the whole solution. This two-step limiter is essential to reduce undesirable effects of the FCT limiter such as accuracy deterioration for smooth solutions, or a frequent switching back and forth between the limited and unlimited schemes. In particular, the numerical experiments highlight a strong improvement of the accuracy of the limited scheme as well as of its ability to capture steady-state solutions that would be otherwise affected when limiting all the DOFs. The former issue has been already observed in the literature [21, 23, 6], where different strategies have been proposed such as bound relaxation [21, 23] or subcell smoothness indicator [34, 6].

We also analyze the inversion of the linear system resulting from the time implicit discretization to be solved at each time step. The linear system is large, non symmetric, sparse with a sparsity pattern containing dense diagonal and sparse off-diagonal blocks of size the number of DOFs per cell. Efficient inversion could be achieved through the use of block sparse direct or iterative linear solvers. Many algorithms require the inversion of the diagonal blocks as in block-preconditionned Krylov solvers [35, 36, 12], block relaxation schemes [46], etc. We here prove that the diagonal blocks are invertible and propose efficient algorithms for their inversion[1]. We again use the nilpotency of the discrete derivative matrix to inverse the diagonal blocks of the 1D scheme. We use the inversion of the 1D diagonal blocks as building blocks for the inversion of diagonal blocks in multiple space dimensions thanks to the tensor product structure of the discretization operators.

The paper is organized as follows. Section 2 introduces some properties of the DGSEM function space associated to Gauss-Lobatto quadrature rules. The 1D DGSEM is introduced and analyzed in section 3, while section 4 focuses on the DGSEM in two space dimensions. The results are assessed by numerical experiments in one and two space dimensions in section 5 and concluding remarks about this work are given in section 6.

2. The DGSEM discretization in space

2.1. The DGSEM function space

The DGSEM discretization consists in defining a discrete weak formulation of problem (1). Although one of the main advantages of the DG method is its ability to handle complex geometries with unstructured grids, we restrict ourselves to rectangular geometries with a structured grid for the sake of simplicity. The space domain Ω is first discretized with a Cartesian grid $\Omega_h \subset \mathbb{R}^d$ with elements κ labeled as $\kappa_i = [x_{i-1/2}, x_{i+1/2}]$ of size $\Delta x_i = x_{i+1/2} - x_{i-1/2} > 0$, $1 \leq i \leq N_x$, for $d = 1$ (see Fig. 1); $\kappa_{ij} = [x_{i-1/2}, x_{i+1/2}] \times [y_{j-1/2}, y_{j+1/2}]$ of size $\Delta x_i \Delta y_j = (x_{i+1/2} - x_{i-1/2})(y_{j+1/2} - y_{j-1/2}) > 0$, $1 \leq i \leq N_x$, $1 \leq j \leq N_y$, for $d = 2$ (see Fig. 1), etc. We also set $h \coloneqq \min_{\kappa \in \Omega_h} |\kappa|^{\frac{1}{d}}$.

The approximate solution to (1) is sought under the form (with some abuse in the notation for the indices and exponents that will be clarified below)

$$u_h(\mathbf{x}, t) = \sum_{k=1}^{N_p} \phi_\kappa^k(\mathbf{x}) U_\kappa^k(t) \quad \forall \mathbf{x} \in \kappa,\ \kappa \in \Omega_h,\ \forall t \geq 0, \tag{4}$$

[1] A repository of the algorithms for block inversion is available at `https://github.com/rueljean/fast_DGSEM_block_inversion`. Consult Appendix B for a description of the repository.

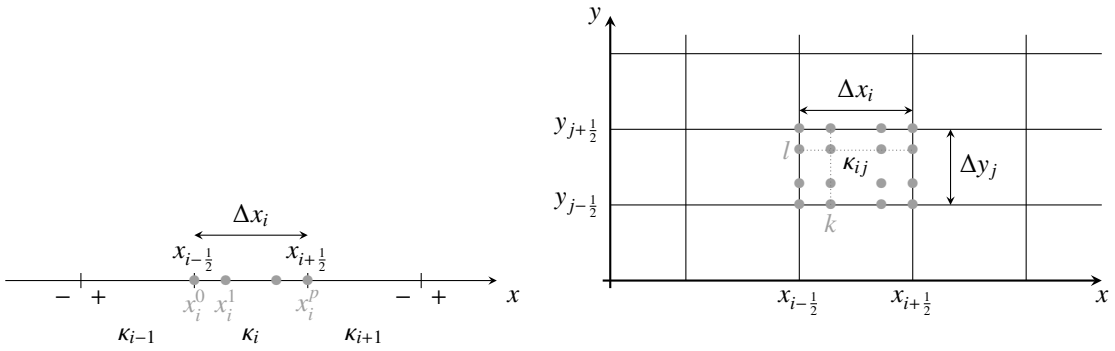


Figure 1: Meshes with positions of quadrature points for $p = 3$ (bullets •): (left) in element κ_i for $d = 1$; (right) in element κ_{ij} for $d = 2$.

where $(U_\kappa^k)_{1\le k\le N_p}$ are the DOFs in the element κ. The subset $(\phi_\kappa^k)_{1\le k\le N_p}$ constitutes a basis of $\mathcal{V}_h^p$, the space of piecewise discrete polynomials of degree p at most, restricted onto the element κ and $N_p = (p+1)^d$ is its dimension. We use tensor product in each space direction of Lagrange interpolation polynomials $(\ell_k)_{0\le k\le p}$ associated to the Gauss-Lobatto quadrature nodes over $I = [-1, 1]$, $\xi_0 = -1 < \xi_1 < \cdots < \xi_p = 1$:

$$\ell_k(\xi) = \prod_{l=0, l\neq k}^{p} \frac{\xi - \xi_l}{\xi_k - \xi_l}, \quad 0 \le k \le p, \tag{5}$$

which satisfy $\ell_k(\xi_l) = \delta_{kl}$, $0 \le k, l \le p$, with δ_{kl} the Kronecker delta.

For the sake of clarity, we now replace the κ index by cell indices in the Cartesian mesh, $i \in \mathbb{N}$ in 1D, $i, j \in \mathbb{N}$ in 2D, while the exponents $0 \le k, l \le p$ refer to the DOFs indices (see Fig. 1). The basis functions are thus defined for $d = 1$ by $\phi_i^k(x) = \ell_k(\frac{2}{\Delta x_i}(x - x_{i-\frac{1}{2}}) - 1)$ and for $d = 2$ by $\phi_{ij}^{kl}(\mathbf{x}) = \ell_k(\frac{2}{\Delta x_i}(x - x_{i-\frac{1}{2}}) - 1)\ell_l(\frac{2}{\Delta y_j}(y - y_{j-\frac{1}{2}}) - 1)$, and so on.

The DGSEM scheme considered in this work uses Gauss-Lobatto quadrature rules to approximate the integrals over elements $\int_{-1}^{1} f(\xi)d\xi \simeq \sum_{k=0}^{p} \omega_k f(\xi_k)$ with $\omega_k > 0$ and $\sum_{k=0}^{p} \omega_k = \int_{-1}^{1} ds = 2$, the weights and ξ_k the nodes over I of the quadrature rule.

2.2. Derivatives of the Lagrange polynomials

It is convenient to introduce the discrete derivative matrix $\mathbf{D}$ [25] with entries D_{kl} defined by

$$D_{kl} = \ell_l'(\xi_k), \quad 0 \le k, l \le p. \tag{6}$$

Note that we have $\ker \mathbf{D} = \mathbb{P}^0(I)$ and by the rank-nullity theorem $\mathbf{D}$ is of rank p. We will also consider $\mathbf{D}^{(\alpha)}$ the generalization to αth-order derivatives:

$$D_{kl}^{(\alpha)} = \ell_l^{(\alpha)}(\xi_k), \quad 0 \le k, l \le p, \quad \alpha \ge 0, \tag{7}$$

with the conventions $D_{kl}^{(0)} = \ell_l(\xi_k) = \delta_{kl}$ and $D_{kl}^{(1)} = \ell_l'(\xi_k) = D_{kl}$. The matrix $\mathbf{D}$ maps any element of $\mathcal{V}_h^p$ to its derivative in $\mathcal{V}_h^{p-1} \subset \mathcal{V}_h^p$ and a direct calculation gives $\mathbf{D}^{(\alpha)} = \mathbf{D}^\alpha$, and since the $(\ell_k)_{0\le k\le p}$ are polynomials of degree p, the matrix $\mathbf{D}$ is nilpotent:

$$\mathbf{D}^{(p+1)} = \mathbf{D}^{p+1} = 0,$$

so one can easily invert the following matrices

$$(\mathbf{I} - y\mathbf{D})^{-1} = \sum_{k=0}^{p} y^k \mathbf{D}^{(k)} \quad \forall y \in \mathbb{R}, \tag{8}$$

which corresponds to the truncated matrix series associated to the Taylor series of the function $x \mapsto (1-yx)^{-1} = \sum_{k\geq 0}(yx)^k$ for $|xy| < 1$. Likewise $\mathbf{D}^{(p)}$ has columns with constant coefficients since $\ell_l^{(p)}$ is a constant function and its entries are easily obtained from (5):

$$D_{kl}^{(p)} = \ell_l^{(p)}(\xi_k) = p! \prod_{m=0, m\neq l}^{p} \frac{1}{\xi_l - \xi_m} \quad \forall 0 \leq k, l \leq p. \tag{9}$$

Integrating $\ell_k^{(\alpha)}$ over I leads to the generalized integration relation

$$\sum_{l=0}^{p} \omega_l D_{lk}^{(\alpha)} = D_{pk}^{(\alpha-1)} - D_{0k}^{(\alpha-1)} \quad \forall 0 \leq k \leq p, \quad \alpha \geq 1, \tag{10}$$

which is the discrete counterpart to $\int_{-1}^{1} \ell_k^{(\alpha)}(\xi)d\xi = \ell_k^{(\alpha-1)}(1) - \ell_k^{(\alpha-1)}(-1)$ and for $\alpha = 1$ we get

$$\sum_{l=0}^{p} \omega_l D_{lk} = \delta_{kp} - \delta_{k0}, \quad 0 \leq k \leq p. \tag{11}$$

Finally, as noticed in [18], the DGSEM satisfies the following important relation known as the summation-by-parts (SBP) property [47] and corresponds to the discrete counterpart to integration by parts:

$$\omega_k D_{kl} + \omega_l D_{lk} = \delta_{kp}\delta_{lp} - \delta_{k0}\delta_{l0}, \quad 0 \leq k, l \leq p. \tag{12}$$

3. Time implicit discretization in one space dimension

We here consider (1) in one space dimension, $d = 1$ and flux cu with $c > 0$, over a unit domain $\Omega = (0, 1)$ and consider periodic conditions $u(0, t) = u(1, t)$ which makes the analysis difficult due to the existence of an upper block in the matrix. This analysis however encompasses the case of more general boundary conditions (see remark 3.3).

3.1. Space-time discretization

The discretization in space of problem (1) is obtained by multiplying (1a) by a test function v_h in $\mathcal{V}_h^p$ where u is replaced by the approximate solution (4), then integrating by parts in space over elements κ_i and replacing the physical fluxes at interfaces by two-point numerical fluxes:

$$\frac{\omega_k \Delta x_i}{2} \partial_t U_i^k + R_i^k(u_h) = 0, \quad 1 \leq i \leq N_x,\ 0 \leq k \leq p, \tag{13a}$$

with

$$R_i^k(u_h) = -\sum_{l=0}^{p} \omega_l D_{lk} f(U_i^l) + \delta_{kp} h(U_i^p, U_{i+1}^0) - \delta_{k0} h(U_{i-1}^p, U_i^0), \tag{13b}$$

where we have used the conventions $U_0^p = U_{N_x}^p$ and $U_{N_x+1}^0 = U_1^0$ to impose the periodic boundary condition. In what follows, we consider the upwind flux. Since $c > 0$, this flux then reads: $h(u^-, u^+) = cu^-$.

We now focus on a time implicit discretization with a backward Euler method in (13a) and the fully discrete scheme reads

$$\frac{\omega_k}{2} U_i^{k,n+1} + \lambda_i \Big(-\sum_{l=0}^{p} \omega_l D_{lk} U_i^{l,n+1} + \delta_{kp} U_i^{p,n+1} - \delta_{k0} U_{i-1}^{p,n+1} \Big) = \frac{\omega_k}{2} U_i^{k,n}, \quad 1 \leq i \leq N_x,\ 0 \leq k \leq p,\ n \geq 0, \tag{14}$$

with $\lambda_i = c\frac{\Delta t^{(n)}}{\Delta x_i}$, $\Delta t^{(n)} = t^{(n+1)} - t^{(n)} > 0$ the time step, with $t^{(0)} = 0$, and using the notations $u_h^{(n)}(\cdot) = u_h(\cdot, t^{(n)})$ and $U_i^{k,n} = U_i^k(t^{(n)})$. Summing (14) over $0 \leq k \leq p$ gives

$$\langle u_h^{(n+1)} \rangle_i + \lambda_i (U_i^{p,n+1} - U_{i-1}^{p,n+1}) = \langle u_h^{(n)} \rangle_i \quad \forall 1 \leq i \leq N_x,\ n \geq 0, \tag{15}$$

for the cell-averaged solution

$$\langle u_h^{(n)} \rangle_i \coloneqq \sum_{k=0}^{p} \frac{\omega_k}{2} U_i^{k,n}. \tag{16}$$

It is convenient to also consider (14) in vector form as

$$\mathbf{M}\mathbf{U}_i^{n+1} = \mathbf{M}\mathbf{U}_i^{n} + \lambda_i \Big((2\mathbf{D}^\top \mathbf{M} - \mathbf{e}_p \mathbf{e}_p^\top)\mathbf{U}_i^{n+1} + \mathbf{e}_0 \mathbf{e}_p^\top \mathbf{U}_{i-1}^{n+1} \Big), \quad 1 \le i \le N_x,\ n \ge 0. \tag{17}$$

where, by $\mathbf{M} = \frac{1}{2}\,\mathrm{diag}(\omega_0, \dots, \omega_p)$ we denote the mass matrix multiplied by $\frac{1}{2}$ with some abuse in the notation, while $(\mathbf{e}_k)_{0\le k\le p}$ is the canonical basis of $\mathbb{R}^{p+1}$ and $\mathbf{U}_i^n = (U_i^{k,n})_{0\le k\le p}$.

Finally, we derive the discrete counterpart to the inequality (2) for the square entropy. Left multiplying (17) by $(\eta'(U_i^{0\le k\le p,n+1}))^\top = \mathbf{U}_i^{(n+1)}$, solutions to (14) satisfy the following inequality for the discrete square entropy $\frac{1}{2}\langle u_h^2 \rangle_i$

$$\frac{1}{2}\langle (u_h^{(n+1)})^2 \rangle_i - \frac{1}{2}\langle (u_h^{(n)})^2 \rangle_i + \frac{\lambda_i}{2}((U_i^{p,n+1})^2 - (U_{i-1}^{p,n+1})^2) \le 0,$$

which brings existence and uniqueness of solutions to (14) in $L^2(\Omega_h \times \cup_{n\ge 0}(t^{(n)}, t^{(n+1)}), \mathbb{R})$.

3.2. The M-matrix framework

Before starting the analysis, we introduce the M-matrix framework that will be useful in the following. We first define the set $\mathcal{Z}^{n\times n}$ of all the $n \times n$ real matrices with nonpositive off-diagonal entries:

$$\mathcal{Z}^{n\times n} = \left\{ \mathbf{A} = (a_{ij}) \in \mathbb{R}^{n\times n} : a_{ij} \le 0, i \ne j \right\}.$$

Different characterizations of M-matrices exist [38] and we use the following definition and characterizations [38]:

Definition 3.1. *A matrix $\mathbf{A} \in \mathcal{Z}^{n\times n}$ is called an M-matrix if $\mathbf{A}$ is inverse-positive. That is $\mathbf{A}^{-1}$ exists and each entry of $\mathbf{A}^{-1}$ is nonnegative.*

Theorem 3.1. *A matrix $\mathbf{A} \in \mathcal{Z}^{n\times n}$ is an M-matrix if and only if $\mathbf{A}$ is semi-positive. That is, there exists $\mathbf{x} = (x_1, \dots, x_n)^\top$ with $x_i > 0$ such that $(\mathbf{A}\mathbf{x})_i > 0$ for all $1 \le i \le n$.*

Theorem 3.2. *A matrix $\mathbf{A} \in \mathcal{Z}^{n\times n}$ is an M-matrix if $\mathbf{A}$ has all positive diagonal elements and it is strictly diagonally dominant, $a_{ii} > \sum_{j\ne i} |a_{ij}|$ for all $1 \le i \le n$.*

M-matrices will be used as a tool to prove positivity preservation for the DGSEM scheme which is equivalent to prove a discrete maximum principle (see lemma 3.1).

3.3. Maximum principle for the cell average

Following [39], we here prove in theorem 3.3 a weaken discrete maximum principle for the cell average, $m \le \langle u_h^{(n+1)} \rangle_i \le M$. We then use the linear scaling limiter from [58] to enforce all the DOFs at time $t^{(n+1)}$ to be in the range $[m, M]$ (see section 5.1). We will use the following result that shows that for the linear and conservative scheme (14), maximum-principle preservation and positivity preservation are equivalent. Note that this result is sufficient, but not necessary as we focus on a maximum principle on the cell averages, not all the DOFs.

Lemma 3.1. *To prove a discrete maximum principle for the DGSEM scheme* (14)*, it is enough to prove that it is positivity preserving.*

Proof. From (11) we obtain

$$\frac{\omega_k}{2} = \frac{\omega_k}{2} + \lambda_i \Big(\sum_{l=0}^{p} \omega_l D_{lk} - \delta_{kp} + \delta_{k0} \Big),$$

and subtracting the above equation multiplied by m defined in (3) from (14), then subtracting (14) from the above equation multiplied by M in (3), we deduce that both $(U_{i\in\mathbb{Z}}^{0\le k\le p,n\ge 0} - m)$ and $(M - U_{i\in\mathbb{Z}}^{0\le k\le p,n\ge 0})$ satisfy (14). As a consequence, the positivity preserving property, $U_{i\in\mathbb{Z}}^{0\le k\le p,n} \ge 0$ implies $U_{i\in\mathbb{Z}}^{0\le k\le p,n+1} \ge 0$, is equivalent to the discrete maximum principle, $m \le U_{i\in\mathbb{Z}}^{0\le k\le p,n} \le M$ implies $m \le U_{i\in\mathbb{Z}}^{0\le k\le p,n+1} \le M$. □

Using (8) with $y = 2\lambda_i$ to invert (17), we get

$$\frac{\omega_k}{2} U_i^{k,n+1} = \sum_{l=0}^{p} \frac{\omega_l}{2} \mathcal{D}_{kl}^i U_i^{l,n} - \lambda_i (\mathcal{D}_{kp}^i U_i^{p,n+1} - \mathcal{D}_{k0}^i U_{i-1}^{p,n+1}),$$

where the $\mathcal{D}_{kl}^i$ denote the entries of the matrix

$$\boldsymbol{\mathcal{D}}_i \coloneqq (\mathbf{I} - 2\lambda_i \mathbf{D}^\top)^{-1} \overset{(8)}{=} \sum_{l=0}^{p} (2\lambda_i \mathbf{D}^\top)^l. \tag{18}$$

We use (15) to get $\lambda_i U_{i-1}^{p,n+1} = \lambda_i U_i^{p,n+1} + \langle u_h^{(n+1)} \rangle_i - \langle u_h^{(n)} \rangle_i$ and injecting this result into the above expression for $U_i^{p,n+1}$ gives

$$\sigma_i^p U_i^{p,n+1} = \xi_i^{p,n} + 2\mathcal{D}_{p0}^i \Big(\langle u_h^{(n+1)} \rangle_i - \langle u_h^{(n)} \rangle_i \Big),$$

where

$$\sigma_i^p = \omega_p + 2\lambda_i (\mathcal{D}_{pp}^i - \mathcal{D}_{p0}^i), \quad \xi_i^{p,n} = \sum_{l=0}^{p} \omega_l \mathcal{D}_{pl}^i U_i^{l,n}. \tag{19}$$

Further using the above expression to eliminate $U_i^{p,n+1}$ and $U_{i-1}^{p,n+1}$ from the cell-averaged scheme (15), we finally obtain

$$\begin{aligned}
\left(1 + 2\lambda_i \frac{\mathcal{D}_{p0}^i}{\sigma_i^p}\right) \langle u_h^{(n+1)} \rangle_i - 2\lambda_i \frac{\mathcal{D}_{p0}^{i-1}}{\sigma_{i-1}^p} \langle u_h^{(n+1)} \rangle_{i-1} &= \left(1 + 2\lambda_i \frac{\mathcal{D}_{p0}^i}{\sigma_i^p}\right) \langle u_h^{(n)} \rangle_i - 2\lambda_i \frac{\mathcal{D}_{p0}^{i-1}}{\sigma_{i-1}^p} \langle u_h^{(n)} \rangle_{i-1} - \lambda_i \left(\frac{\xi_i^{p,n}}{\sigma_i^p} - \frac{\xi_{i-1}^{p,n}}{\sigma_{i-1}^p} \right) \\
&= \langle u_h^{(n)} \rangle_i - \lambda_i \frac{\xi_i^{p,n} - 2\mathcal{D}_{p0}^i \langle u_h^{(n)} \rangle_i}{\sigma_i^p} + \lambda_i \frac{\xi_{i-1}^{p,n} - 2\mathcal{D}_{p0}^{i-1} \langle u_h^{(n)} \rangle_{i-1}}{\sigma_{i-1}^p} \\
&= \sum_{k=0}^{p} \frac{\omega_k}{2} \left(\left(1 - \frac{2\lambda_i \left(\mathcal{D}_{pk}^i - \mathcal{D}_{p0}^i \right)}{\sigma_i^p} \right) U_i^{k,n} + \left(\frac{2\lambda_i \left(\mathcal{D}_{pk}^{i-1} - \mathcal{D}_{p0}^{i-1} \right)}{\sigma_{i-1}^p} \right) U_{i-1}^{k,n} \right),
\end{aligned} \tag{20}$$

where we have used (16) and (19) in the last step.

Let us now derive conditions on λ_i for which the above relation preserves a discrete maximum principle for the cell-averaged solution. According to lemma 3.1, it is enough to prove that the scheme preserves positivity, i.e., $U_{1 \le i \le N_x}^{0 \le k \le p,n} \ge 0$ imply $\langle u_h^{(n+1)} \rangle_{1 \le i \le N_x} \ge 0$. We will thus show that, under some conditions on λ_i, the matrix stemming from the linear system (20) for the $\langle u_h^{(n+1)} \rangle_{1 \le j \le N_x}$ is an M-matrix by using the characterization in theorem 3.1 and that its RHS is a nonnegative combination of the $U_i^{k,n}$ and $U_{i-1}^{k,n}$. Assuming the DOFs at time $t^{(n)}$ are in the range $[m, M]$, so will do the cell-averaged solutions $\langle u_h^{(n+1)} \rangle_{1 \le i \le N_x}$ according to definition 3.1.

In view of (20), conditions for the RHS to be nonnegative read

$$\sigma_i^p - 2\lambda_i (\mathcal{D}_{pk}^i - \mathcal{D}_{p0}^i) = \omega_p + 2\lambda_i (\mathcal{D}_{pp}^i - \mathcal{D}_{pk}^i) \ge 0, \quad \mathcal{D}_{pk}^i - \mathcal{D}_{p0}^i \ge 0 \quad \forall 0 \le k \le p,\ 1 \le i \le N_x, \tag{21}$$

with $\sigma_i^p > 0$, while we impose the off-diagonal entries to be negative through

$$\sigma_i^p = \omega_p + 2\lambda_i (\mathcal{D}_{pp}^i - \mathcal{D}_{p0}^i) > 0, \quad \mathcal{D}_{p0}^i \ge 0. \tag{22}$$

The strict inequality on the $\mathcal{D}_{p0}^i$ allows to satisfy theorem 3.1 by choosing the vector $\mathbf{x}$ such that $x_i = \prod_{j=1, j \ne i}^{N_x} \frac{\mathcal{D}_{p0}^j}{\sigma_j^p} > 0$ for all $1 \le i \le N_x$ and we obtain $(\mathbf{A}\mathbf{x})_i = x_i > 0$ from (22).

Lemma 3.2. *For all $p \ge 1$, there exists a finite $\lambda_{min} = \lambda_{min}(p) \ge 0$ such that conditions* (21) *and* (22) *are satisfied for all $\lambda_i > \lambda_{min}$, $1 \le i \le N_x$.*

Table 1: Lower bounds on the non-dimensional time step $\lambda_i > \lambda_{min}$, $1 \le i \le N_x$, for (21) and (22) to hold, which make (20) maximum principle preserving.

p	1	2	3	4	5	6
λ_{min}	0	$\frac{1}{4}$	$\frac{1+\sqrt{5}}{6(5-\sqrt{5})}$	0.150346	0.147568	0.109977

Proof. Let us consider the first condition in (21), similar arguments hold for all other conditions. For a fixed $0 \le k \le p$, use (18) to rewrite

$$\mathcal{D}^i_{pp} - \mathcal{D}^i_{pk} = \sum_{l=0}^{p} (2\lambda_i)^l \left(D^{(l)}_{pp} - D^{(l)}_{kp}\right) = \sum_{l=0}^{p-1} (2\lambda_i)^l \left(D^{(l)}_{pp} - D^{(l)}_{kp}\right),$$

since by (9), we have $D^{(p)}_{kp} = D^{(p)}_{pp}$. Hence for large λ_i, we have

$$\mathcal{D}^i_{pp} - \mathcal{D}^i_{pk} \underset{\lambda_i}{\sim} (2\lambda_i)^{p-1} \left(D^{(p-1)}_{pp} - D^{(p-1)}_{kp}\right)$$

and we are going to show that this is a positive quantity. By using the linearity of $\ell^{(p-1)}_p(\cdot)$, we have $D^{(p-1)}_{kp} = \frac{1-\xi_k}{2} D^{(p-1)}_{0p} + \frac{1+\xi_k}{2} D^{(p-1)}_{pp}$. Then, we obtain

$$D^{(p-1)}_{pp} - D^{(p-1)}_{kp} = \frac{1-\xi_k}{2}\left(D^{(p-1)}_{pp} - D^{(p-1)}_{0p}\right) \overset{(10)}{=} \frac{1-\xi_k}{2} \sum_{l=0}^{p} \omega_l D^{(p)}_{lp} \overset{(9)}{=} (1-\xi_k) D^{(p)}_{pp} \overset{(9)}{=} \frac{(1-\xi_k)p!}{\prod_{l=0}^{p-1}(1-\xi_l)} > 0,$$

which concludes the proof. □

The following theorem immediately follows.

Theorem 3.3. *Under the conditions $\lambda_{1\le i\le N_x} > \lambda_{min}$ defined in lemma 3.2, the DGSEM scheme* (14) *is maximum principle preserving for the cell-averaged solution:*

$$m \le U^{k,n}_i \le M \quad \forall 1 \le i \le N_x,\ 0 \le k \le p \quad \Rightarrow \quad m \le \langle u^{(n+1}_h \rangle_i \le M \quad \forall 1 \le i \le N_x.$$

Tab. 1 indicates the lower bounds on the λ_i as a function of the polynomial degree p evaluated from the conditions (21) and (22). We observe that the second-order in space scheme, $p = 1$, is unconditionally maximum principle preserving, while the lower bound decreases with increasing p values for $p \ge 2$. These bounds are different from those obtained in [39, Tab. 1] for the modal DG scheme with Legendre polynomials as function basis. In particular, the modal DG scheme with $p = 1$ is not unconditionally maximum principle preserving and is seen to require a larger CFL value for larger p values.

Remark 3.1 (Linear hyperbolic systems). The above results also apply to the case of linear hyperbolic systems of size n_{eq} with constant coefficients $\partial_t \mathbf{u} + \mathbf{A}\partial_x \mathbf{u} = 0$ with $\mathbf{A}$ diagonalizable in $\mathbb{R}$ with eigenvalues ψ_k, normalized left and right eigenvectors $\mathbf{l}_k$ and $\mathbf{r}_k$ such that $\mathbf{l}^\top_k \mathbf{r}_l = \delta_{kl}$. Assuming that the right eigenvectors form a basis of $\mathbb{R}^{n_{eq}}$ and setting $\mathbf{u} = \sum_k u_k \mathbf{r}_k$, each component satisfies a maximum principle: $\partial_t u_k + \psi_k \partial_x u_k = 0$. Using a Roe flux, $\mathbf{h}(\mathbf{u}^-, \mathbf{u}^+) = \frac{1}{2}\mathbf{A}(\mathbf{u}^- + \mathbf{u}^+) + \frac{1}{2}\sum_k |\psi_k|(u^-_k - u^+_k)\mathbf{r}_k$, the time implicit DGSEM decouples into n_{eq} independent schemes (14) for u_k upon left multiplication by $\mathbf{l}_k$ since $\mathbf{l}^\top_k \mathbf{h}(\mathbf{u}^-, \mathbf{u}^+) = \frac{\psi_k}{2}(u^-_k + u^+_k) + \frac{|\psi_k|}{2}(u^-_k - u^+_k)$ reduces to the upwind flux.

Remark 3.2 (Geometric source). Theorem 3.3 with the bounds from lemma 3.2 also applies to linear equations with a geometric source term:

$$\partial_t u + c\partial_x u = s(x) \quad \text{in } \Omega \times (0, \infty),$$

with $s(\cdot) \ge 0$. Providing nonnegative initial and boundary data are imposed, the entropy solution remains nonnegative for all time. The DGSEM scheme (14) for the discretization of the above equation remains the same by substituting $U^{k,n}_i + s(x^k_i)\Delta t$ for $U^{k,n}_i$ in the RHS. The conditions to obtain an M-matrix in (20) are therefore unchanged and only the RHS is modified by the above change of variable on $U^{k,n}_i$. Adding a source term, lemma 3.1 no longer holds, but the present DGSEM is positivity-preserving for the cell-averaged solution under the same conditions as in theorem 3.3.

3.4. Linear system solution

Equations (14) or (17) result in a banded block linear system

$$\mathbb{A}_{1d}\mathbf{U}^{(n+1)} = \mathbb{M}_{1d}\mathbf{U}^{(n)} \tag{23}$$

of size $N_x(p+1)$ with blocks of size $p+1$ to be solved for $\mathbf{U}^{(n+1)}$ where $\mathbf{U}_{1+k+(p+1)i} = U_i^k$ and $\mathbb{M}_{1d}$ is the global mass matrix. Using the block structure of (23) is usually important for its efficient resolution with either direct or iterative block-based solvers. We will also propose a direct algorithm below based on the inversion of the diagonal block only. In most cases, we need to invert the diagonal blocks and we now prove that they are unconditionally invertible, while a fast algorithm for the inversion of (23) is proposed in Appendix C.

Lemma 3.3. *For all $p \geq 1$, the diagonal blocks $\mathbf{L}_{1d}\mathbf{M}$ of $\mathbb{A}_{1d}$ in the linear system* (17)*, with*

$$\mathbf{L}_{1d} = \mathbf{I} - 2\lambda\boldsymbol{\mathcal{L}}, \quad \boldsymbol{\mathcal{L}} = \mathbf{D}^\top - \frac{1}{\omega_p}\mathbf{e}_p\mathbf{e}_p^\top, \tag{24}$$

are invertible for any $\lambda > 0$.

Proof. Let us prove that $\mathbf{L}_{1d}$ is invertible: assume that $\mathbf{L}_{1d}\mathbf{u} = 0$ for some $\mathbf{u} = (u_0, \dots, u_p)^\top$, then by (24) we have

$$(\mathbf{I} - 2\lambda\mathbf{D}^\top)\mathbf{u} = -\frac{2\lambda}{\omega_p}\mathbf{e}_p\mathbf{e}_p^\top\mathbf{u} = -\frac{2\lambda u_p}{\omega_p}\mathbf{e}_p \quad \Rightarrow \quad \mathbf{u} = -\frac{2\lambda u_p}{\omega_p}\boldsymbol{\mathcal{D}}\mathbf{e}_p,$$

with $\boldsymbol{\mathcal{D}} = (\mathbf{I} - 2\lambda\mathbf{D}^\top)^{-1}$ given by (18). Hence $u_k = -\frac{2\lambda u_p}{\omega_p}\mathcal{D}_{kp}$ and for $k = p$ we get $(\omega_p + 2\lambda\mathcal{D}_{pp})u_p = 0$, so $u_p = 0$ since $\omega_p + 2\lambda\mathcal{D}_{pp} > 0$ from (22) and we conclude that $\mathbf{u} = 0$. Note that (24) is invertible for all $\lambda > 0$ since we have $\mathcal{D}_{pp} = 1 + \sum_{l=1}^{p}(2\lambda)^l D_{pp}^{(l)} > 0$. Indeed, by differentiating (5) l-times we obtain

$$D_{pp}^{(l)} = \ell_p^{(l)}(1) = \sum_{k_1=0}^{p}\sum_{k_2=0,k_2\neq k_1}^{p}\cdots\sum_{k_l=1,k_l\notin\{k_1,\dots,k_{l-1}\}}^{p}\prod_{m=1}^{l}\frac{1}{1-\xi_{k_m}} > 0 \quad \forall 1 \leq l \leq p.$$

□

Remark 3.3 (Dirichlet boundary condition). The case of an inflow boundary condition, $u(0,t) = g(t) \in [m, M]$, results in a similar linear system (23) with the only difference that $U_{-1}^{p,n+1} = g(t^{(n+1)})$ in (13b). As a consequence, $\mathbb{A}_{1d} = \mathbb{A}_0$ in (23), where $\mathbb{A}_0$ is a block lower triangular matrix with diagonal blocks $\mathbf{ML}_{1d}$, subdiagonal blocks $-\lambda_i\mathbf{e}_0\mathbf{e}_p^\top$, and the $\mathbf{L}_{1d}$ defined in (24). The system (23) is therefore easily solved by block forward substitution since the diagonal blocks are invertible. Likewise, the cell-averaged solution is maximum principle preserving under the same conditions in lemma 3.2 as with periodic boundary conditions. Indeed, similar manipulations as for the derivation of (20) give for the first cell $i = 1$

$$(\omega_p + 2\lambda_1\mathcal{D}_{pp}^1)\langle u_h^{(n+1)}\rangle_1 = \sum_{k=0}^{p}\frac{\omega_k}{2}\Big(\big(\omega_p + 2\lambda_1(\mathcal{D}_{pp}^1 - \mathcal{D}_{pk}^1)\big)U_1^{k,n} + \big(\omega_p + 2\lambda_1(\mathcal{D}_{pp}^1 - \mathcal{D}_{p0}^1)\big)g(t^{(n+1)})\Big),$$

and we conclude from (21) and (22).

4. Time implicit discretization in two space dimensions

We now consider a 2D linear problem with constant coefficients:

$$\partial_t u + c_x\partial_x u + c_y\partial_y u = 0, \quad \text{in } \Omega\times(0,\infty), \tag{25a}$$

$$u(\cdot,0) = u_0(\cdot), \quad \text{in } \Omega, \tag{25b}$$

with boundary conditions on $\partial\Omega$ and we again assume $c_x \geq 0$ and $c_y \geq 0$ without loss of generality. We also assume $\Omega = \mathbb{R}^2$ for the analysis, which amounts in practice to consider a rectangular domain with periodic boundary conditions. As in the 1D case, considering inflow and outflow boundary conditions results in a block lower triangular system to be solved and hence an easier analysis. The results in this section may be easily generalized to three space dimensions.

4.1. Space-time discretization

We consider a Cartesian mesh with rectangular elements of measure $|\kappa_{ij}| = \Delta x_i \times \Delta y_j$ for all i, j in $\mathbb{Z}$. Using again a time implicit discretization with a backward Euler method and upwind numerical fluxes, the fully discrete scheme reads

$$\begin{aligned}\frac{\omega_k\omega_l}{4}(U_{ij}^{kl,n+1} - U_{ij}^{kl,n}) - \frac{\omega_l}{2}\lambda_{x_i}\Big(\sum_{m=0}^{p}\omega_m D_{mk}U_{ij}^{ml,n+1} - \delta_{kp}U_{ij}^{pl,n+1} + \delta_{k0}U_{(i-1)j}^{pl,n+1}\Big)\\ -\frac{\omega_k}{2}\lambda_{y_j}\Big(\sum_{m=0}^{p}\omega_m D_{ml}U_{ij}^{km,n+1} - \delta_{lp}U_{ij}^{kp,n+1} + \delta_{l0}U_{i(j-1)}^{kp,n+1}\Big) = 0,\end{aligned} \tag{26}$$

where $\lambda_{x_i} = \frac{c_x\Delta t}{\Delta x_i}$ and $\lambda_{y_j} = \frac{c_y\Delta t}{\Delta y_j}$. We again use the conventions $U_{0j}^{pl} = U_{N_x j}^{pl}$ and $U_{i0}^{kp} = U_{iN_y}^{kp}$ to take the periodic boundary conditions into account. Using a vector storage of the DOFs as $(\mathbf{U}_{ij})_{n_{kl}} = U_{ij}^{kl}$ with $1 \le n_{kl} := 1+k+l(p+1) \le N_p$ and $N_p = (p+1)^2$, it will be convenient to rewrite the scheme under vector form as

$$\begin{aligned}(\mathbf{M}\otimes\mathbf{M})(\mathbf{U}_{ij}^{n+1} - \mathbf{U}_{ij}^{n}) - \lambda_{x_i}\Big(\mathbf{M}\otimes(2\mathbf{D}^\top\mathbf{M} - \mathbf{e}_p\mathbf{e}_p^\top)\Big)\mathbf{U}_{ij}^{n+1} - \lambda_{x_i}(\mathbf{M}\otimes\mathbf{e}_0\mathbf{e}_p^\top)\mathbf{U}_{(i-1)j}^{n+1}\\ - \lambda_{y_j}\Big((2\mathbf{D}^\top\mathbf{M} - \mathbf{e}_p\mathbf{e}_p^\top)\otimes\mathbf{M}\Big)\mathbf{U}_{ij}^{n+1} - \lambda_{y_j}(\mathbf{e}_0\mathbf{e}_p^\top\otimes\mathbf{M})\mathbf{U}_{i(j-1)}^{n+1} = 0,\end{aligned} \tag{27}$$

where $\mathbf{M} = \frac{1}{2}\,\mathrm{diag}(\omega_0,\dots,\omega_p)$, $(\mathbf{e}_k)_{0\le k\le p}$ is the canonical basis of $\mathbb{R}^{p+1}$, and $\otimes$ denotes the Kronecker product [51, 50]: $(\mathbf{A}\otimes\mathbf{B})_{n_{kl}n_{k'l'}} = \mathbf{A}_{ll'}\mathbf{B}_{kk'}$, which satisfies $(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = \mathbf{AC}\otimes\mathbf{BD}$, $(\mathbf{A}\otimes\mathbf{B})^{-1} = \mathbf{A}^{-1}\otimes\mathbf{B}^{-1}$, and $(\mathbf{A}\otimes\mathbf{B})^\top = \mathbf{A}^\top\otimes\mathbf{B}^\top$. Likewise, for diagonalizable matrices $\mathbf{A} = \mathbf{R}_A\mathbf{\Psi}_A\mathbf{R}_A^{-1}$ and $\mathbf{B} = \mathbf{R}_B\mathbf{\Psi}_B\mathbf{R}_B^{-1}$, the product $\mathbf{A}\otimes\mathbf{B}$ is also diagonalizable with eigenvalues being the product of eigenvalues of $\mathbf{A}$ and $\mathbf{B}$: $\mathbf{A}\otimes\mathbf{B} = (\mathbf{R}_A\otimes\mathbf{R}_B)(\mathbf{\Psi}_A\otimes\mathbf{\Psi}_B)(\mathbf{R}_A\otimes\mathbf{R}_B)^{-1}$.

Summing (26) over $0 \le k, l \le p$ gives:

$$\langle u_h^{(n+1)}\rangle_{ij} - \langle u_h^{(n)}\rangle_{ij} + \frac{\lambda_{x_i}}{2}\sum_{l=0}^{p}\omega_l\Big(U_{ij}^{pl,n+1} - U_{(i-1)j}^{pl,n+1}\Big) + \frac{\lambda_{y_j}}{2}\sum_{k=0}^{p}\omega_k\Big(U_{ij}^{kp,n+1} - U_{i(j-1)}^{kp,n+1}\Big) = 0, \tag{28}$$

where the cell-average operator reads

$$\langle u_h\rangle_{ij} = \sum_{k=0}^{p}\sum_{l=0}^{p}\frac{\omega_k\omega_l}{4}U_{ij}^{kl}. \tag{29}$$

Finally, left-multiplying (27) by $\mathbf{U}_{ij}^{(n+1)}$ brings L^2 stability:

$$\frac{1}{2}\langle(u_h^{(n+1)})^2\rangle_{ij} - \frac{1}{2}\langle(u_h^{(n)})^2\rangle_{ij} + \frac{\lambda_{x_i}}{2}\sum_{l=0}^{p}\frac{\omega_l}{2}((U_{ij}^{pl,n+1})^2 - (U_{(i-1)j}^{pl,n+1})^2) + \frac{\lambda_{y_j}}{2}\sum_{k=0}^{p}\frac{\omega_k}{2}((U_{ij}^{kp,n+1})^2 - (U_{i(j-1)}^{kp,n+1})^2) \le 0.$$

Note that the 2D discrete difference matrix reads $\mathbf{D}_{2d}^\top = \lambda_{x_i}\mathbf{I}\otimes\mathbf{D}^\top + \lambda_{y_j}\mathbf{D}^\top\otimes\mathbf{I}$, and it may be easily checked that is is also nilpotent, $\mathbf{D}_{2d}^{2p+1} = 0$, so $\mathbf{I} - \mathbf{D}_{2d}$ may be inverted as in the 1D case. Unfortunately, the scheme (27) is in general not maximum principle preserving for the cell average as may be observed in the numerical experiments of section 5. We now propose to modify the scheme by adding graph viscosity to make it maximum principle preserving.

4.2. Maximum principle through graph viscosity

We add a graph viscosity [22] term $\mathbf{V}_{ij}^{(n+1)}$ to the LHS of (27) which becomes

$$\begin{aligned}(\mathbf{M}\otimes\mathbf{M})(\mathbf{U}_{ij}^{n+1} - \mathbf{U}_{ij}^{n}) - \lambda_{x_i}(\mathbf{M}\otimes(2\mathbf{D}^\top\mathbf{M} - \mathbf{e}_p\mathbf{e}_p^\top))\mathbf{U}_{ij}^{n+1} - \lambda_{x_i}(\mathbf{M}\otimes\mathbf{e}_0\mathbf{e}_p^\top)\mathbf{U}_{(i-1)j}^{n+1}\\ - \lambda_{y_j}((2\mathbf{D}^\top\mathbf{M} - \mathbf{e}_p\mathbf{e}_p^\top)\otimes\mathbf{M})\mathbf{U}_{ij}^{n+1} - \lambda_{y_j}(\mathbf{e}_0\mathbf{e}_p^\top\otimes\mathbf{M})\mathbf{U}_{i(j-1)}^{n+1} + \mathbf{V}_{ij}^{(n+1)} = 0,\end{aligned} \tag{30}$$

where

$$\begin{aligned}\mathbf{V}_{ij}^{(n+1)} &= 2d_{ij}\Big(\lambda_{x_i}\mathbf{M}\otimes(\mathbf{M} - \boldsymbol{\omega}\mathbf{1}^\top\mathbf{M}) + \lambda_{y_j}(\mathbf{M} - \boldsymbol{\omega}\mathbf{1}^\top\mathbf{M})\otimes\mathbf{M}\Big)\mathbf{U}_{ij}^{(n+1)}\\ &= 2d_{ij}\Big((\lambda_{x_i}+\lambda_{y_j})\mathbf{I}\otimes\mathbf{I} - \lambda_{x_i}(\mathbf{I}\otimes\boldsymbol{\omega}\mathbf{1}^\top) - \lambda_{y_j}(\boldsymbol{\omega}\mathbf{1}^\top\otimes\mathbf{I})\Big)(\mathbf{M}\otimes\mathbf{M})\mathbf{U}_{ij}^{(n+1)},\end{aligned} \tag{31}$$

with $d_{ij} \geq 0$, $\boldsymbol{\omega} = \frac{1}{2}(\omega_0, \dots, \omega_p)^\top$ and $\mathbf{1} = (1, \dots, 1)^\top \in \mathbb{R}^{p+1}$, which reads componentwise as

$$V_{ij}^{kl,n+1} = d_{ij}\frac{\omega_k\omega_l}{2}\Big(\lambda_{x_i}\sum_{m=0}^{p}\frac{\omega_m}{2}(U_{ij}^{kl,n+1} - U_{ij}^{ml,n+1}) + \lambda_{y_j}\sum_{m=0}^{p}\frac{\omega_m}{2}(U_{ij}^{kl,n+1} - U_{ij}^{km,n+1})\Big). \tag{32}$$

This term keeps conservation of the scheme: $\sum_{k,l} V_{ij}^{kl,n+1} = 0$, so the cell-averaged scheme still satisfies (28). It also enforces the L^2 stability since

$$\begin{aligned}
\mathbf{U}_{ij}\cdot\mathbf{V}_{ij} &\overset{(32)}{=} d_{ij}\sum_{k,l=0}^{p}\frac{\omega_k\omega_l}{2}U_{ij}^{kl}\left(\lambda_{x_i}\sum_{m=0}^{p}\frac{\omega_m}{2}(U_{ij}^{kl} - U_{ij}^{ml}) + \lambda_{y_j}\sum_{m=0}^{p}\frac{\omega_m}{2}(U_{ij}^{kl} - U_{ij}^{km})\right)\\
&= d_{ij}\sum_{k,l=0}^{p}\frac{\omega_k\omega_l}{2}\left(\lambda_{x_i}\sum_{m=0}^{p}\frac{\omega_m}{2}\frac{(U_{ij}^{kl} - U_{ij}^{ml})^2 + (U_{ij}^{kl})^2 - (U_{ij}^{ml})^2}{2} + \lambda_{y_j}\sum_{m=0}^{p}\frac{\omega_m}{2}\frac{(U_{ij}^{kl} - U_{ij}^{km})^2 + (U_{ij}^{kl})^2 - (U_{ij}^{km})^2}{2}\right)\\
&= d_{ij}\sum_{k,l=0}^{p}\frac{\omega_k\omega_l}{2}\left(\lambda_{x_i}\sum_{m=0}^{p}\frac{\omega_m}{2}\frac{(U_{ij}^{kl} - U_{ij}^{ml})^2}{2} + \lambda_{y_j}\sum_{m=0}^{p}\frac{\omega_m}{2}\frac{(U_{ij}^{kl} - U_{ij}^{km})^2}{2}\right) \geq 0.
\end{aligned}$$

We now look for conditions on the linear system (30) to correspond to an M-matrix, thus imposing a maximum principle for the DOFs.

Lemma 4.1. *Under the condition*

$$d_{ij} \geq 2\max_{0\leq k\neq m\leq p}\Big(-\frac{D_{mk}}{\omega_k}\Big), \tag{33}$$

the linear system (30) *is maximum principle preserving.*

Proof. This is a direct application of theorem 3.2 to show that the linear system (30) is defined from an M-matrix. Positivity preservation is then enough to get maximum principle preservation. We rewrite (30) componentwise as

$$a_{kl}U_{ij}^{kl,n+1} + \sum_{m=0,m\neq k}^{p} b_{klm}U_{ij}^{ml,n+1} + \sum_{m=0,m\neq l}^{p} c_{klm}U_{ij}^{km,n+1} - \lambda_{x_i}\frac{\omega_l}{2}\delta_{k0}U_{(i-1)j}^{pl,n+1} - \lambda_{y_j}\frac{\omega_k}{2}\delta_{l0}U_{i(j-1)}^{kp,n+1} = \frac{\omega_k\omega_l}{4}U_{ij}^{kl,n}$$

where the diagonal coefficients read

$$a_{kl} = \frac{\omega_k\omega_l}{4} - \lambda_{x_i}\frac{\omega_l}{2}(\omega_k D_{kk} - \delta_{kp}) - \lambda_{y_j}\frac{\omega_k}{2}(\omega_l D_{ll} - \delta_{lp}) + \frac{\omega_k\omega_l}{2}d_{ij}\Big(\lambda_{x_i}\sum_{m=0,m\neq k}^{p}\frac{\omega_m}{2} + \lambda_{y_j}\sum_{m=0,m\neq l}^{p}\frac{\omega_m}{2}\Big)$$

and are positive since $-(\omega_k D_{kk} - \delta_{kp}) = \frac{1}{2}(\delta_{kp} + \delta_{k0})$ and $-(\omega_l D_{ll} - \delta_{lp}) = \frac{1}{2}(\delta_{lp} + \delta_{l0})$ from the SBP property (12). Likewise, $b_{klm} = -\lambda_{x_i}\frac{\omega_l}{2}\omega_m(D_{mk} + \frac{\omega_k}{2}d_{ij})$ and $c_{klm} = -\lambda_{y_j}\frac{\omega_k}{2}\omega_m(D_{ml} + \frac{\omega_l}{2}d_{ij})$ are nonpositive under (33).

Finally, strict diagonal dominance reads $a_{kl} > -\sum_{m\neq k} b_{klm} - \sum_{m\neq l} c_{klm} + \lambda_{x_i}\frac{\omega_l}{2}\delta_{k0} + \lambda_{y_j}\frac{\omega_k}{2}\delta_{l0}$ since $b_{klm} \leq 0$ and $c_{klm} \leq 0$. This reduces to

$$\frac{\omega_k\omega_l}{4} - \lambda_{x_i}\frac{\omega_l}{2}(\omega_k D_{kk} - \delta_{kp}) - \lambda_{y_j}\frac{\omega_k}{2}(\omega_l D_{ll} - \delta_{lp}) > \lambda_{x_i}\frac{\omega_l}{2}\sum_{m=0,m\neq k}^{p}\omega_m D_{mk} + \lambda_{y_j}\frac{\omega_k}{2}\sum_{m=0,m\neq l}^{p}\omega_m D_{ml} + \lambda_{x_i}\frac{\omega_l}{2}\delta_{k0} + \lambda_{y_j}\frac{\omega_k}{2}\delta_{l0},$$

where all the coefficients from the graph viscosity cancel each other out from (32) and have been removed. This can be rearranged into

$$\frac{\omega_k\omega_l}{4} > \lambda_{x_i}\frac{\omega_l}{2}\Big(\sum_{m=0}^{p}\omega_m D_{mk} - \delta_{kp} + \delta_{k0}\Big) + \lambda_{y_j}\frac{\omega_k}{2}\Big(\sum_{m=0}^{p}\omega_m D_{ml} - \delta_{lp} + \delta_{l0}\Big) \overset{(11)}{=} 0,$$

which is always satisfied and concludes the proof. □

Table 2: Lower bounds (33) on the coefficient d_{ij} for (30) to be maximum-principle preserving.

p	1	2	3	4	5	6
$2 \max_{0\le k\neq m\le p}(-\frac{D_{mk}}{\omega_k})$	1	3	$3(1+\sqrt{5})$	24.8	53.6	102.6

The modified DGSEM scheme with graph viscosity therefore satisfies the maximum principle for large enough d_{ij} values. Table 2 gives the minimum d_{ij} values (33) in lemma 4.1 guaranteeing a maximum principle. Likewise, the diagonal blocks in (30) are now strictly diagonally dominant and hence invertible.

The scheme is however first order in space when $d_{ij} > 0$ and is not used in practice. In the following, it is combined with the high-order scheme within the FCT limiter framework to keep high-order accuracy.

Remark 4.1 (sparse discrete derivative matrix). The accuracy of the low-order DGSEM scheme (30) with graph viscosity may be improved by replacing the discrete derivative matrix (6) with the low-order sparse derivative matrix $\tilde{\mathbf{D}}$ from [34, Sec. 3] defined by

$$\omega_k \tilde{D}_{k(k+1)} = -\omega_k \tilde{D}_{k(k-1)} = \frac{1}{2}, \quad \omega_p \tilde{D}_{pp} = -\omega_0 \tilde{D}_{00} = \frac{1}{2}. \tag{34}$$

The matrix $\tilde{\mathbf{D}}$ indeed still satisfies (11) and (12), so the scheme satisfies the maximum principle, but for a lower dissipation coefficient and applied to a sparser stencil $\max(k-1,0) \le l \le \min(k+1,0)$ instead of $0 \le l \le p$, whence $\omega_k \omega_{k\pm1} d_{ij} = 1$ is enough. Though this modification reduces the dissipation necessary to achieve a maximum principle, the scheme is still first-order accurate and the limiting strategy introduced in section 4.3 is seen to be weakly affected by the amount of dissipation as illustrated in Appendix A (using either $\mathbf{D}$ or $\tilde{\mathbf{D}}$ in the low-order scheme did not change all the numerical results of section 5.3).

4.3. Flux-corrected transport limiter

Following [20], the Flux-Corrected Transport (FCT) limiter [5, 57] can be applied to guarantee a maximum principle by combining the high-order (HO) DGSEM scheme (27) and the low-order (LO) modified DGSEM scheme (30) with graph viscosity. We here propose to use the FCT limiter to guarantee a maximum principle on the cell-averaged solution (29), the maximum principle on all DOFs within the elements being ensured through the use of the linear scaling limiter (see section 5.1) as in one space dimension. This two-step limiting process has been introduced to avoid undesirable effects on the approximate solution when the FCT limiting is applied to all DOFs, such as accuracy deterioration for smooth solutions, or a frequent switching back and forth between the limited and unlimited schemes (see Appendix A for a numerical illustration).

By $u_{h,LO}^{(n+1)}$ (resp., $u_{h,HO}^{(n+1)}$) we denote the solution to the LO scheme (30) (resp., HO scheme (27)). Both are solutions to the cell-averaged scheme (28). Subtracting the cell-averaged for the LO solution from the one for the HO solution gives

$$\begin{aligned}
\langle u_{h,HO}^{(n+1)}\rangle_{ij} - \langle u_{h,LO}^{(n+1)}\rangle_{ij} &\overset{(28)}{=} \lambda_{x_i} \sum_{l=0}^{p} \frac{\omega_l}{2}\Big(U_{(i-1)j,HO}^{pl,n+1} - U_{ij,HO}^{pl,n+1} + U_{ij,LO}^{pl,n+1} - U_{(i-1)j,LO}^{pl,n+1}\Big) \\
&\quad + \lambda_{y_j} \sum_{k=0}^{p} \frac{\omega_k}{2}\Big(U_{i(j-1),HO}^{kp,n+1} - U_{ij,HO}^{kp,n+1} + U_{ij,LO}^{kp,n+1} - U_{i(j-1),LO}^{kp,n+1}\Big) \\
&= \lambda_{x_i} \sum_{l=0}^{p} \frac{\omega_l}{2}\Big(U_{(i-1)j,HO}^{pl,n+1} - U_{(i-1)j,LO}^{pl,n+1}\Big) + \lambda_{x_i} \sum_{l=0}^{p} \frac{\omega_l}{2}\Big(-U_{ij,HO}^{pl,n+1} + U_{ij,LO}^{pl,n+1}\Big) \\
&\quad + \lambda_{y_j} \sum_{k=0}^{p} \frac{\omega_k}{2}\Big(U_{i(j-1),HO}^{kp,n+1} - U_{i(j-1),LO}^{kp,n+1}\Big) + \lambda_{y_j} \sum_{k=0}^{p} \frac{\omega_k}{2}\Big(-U_{ij,HO}^{kp,n+1} + U_{ij,LO}^{kp,n+1}\Big) \\
&=: A_{ij}^{(i-1)j} + A_{ij}^{(i+1)j} + A_{ij}^{i(j-1)} + A_{ij}^{i(j+1)} = \sum_{(r,s)\in\mathcal{S}(i,j)} A_{ij}^{rs}
\end{aligned}$$

with $\mathcal{S}(i,j) = \{(i-1,j);(i+1,j);(i,j-1);(i,j+1)\}$. Note that we have $A_{ij}^{(i-1)j} = -A_{(i-1)j}^{ij}$ and $A_{ij}^{i(j-1)} = -A_{i(j-1)}^{ij}$. Again following [20, Sec. 5.3], we introduce the limiter coefficients defined by

$$P_{ij}^- = \sum_{(r,s)\in\mathcal{S}(i,j)} \min\left(A_{ij}^{rs}, 0\right) \le 0, \qquad Q_{ij}^- = m - \langle u_{LO}^{(n+1)}\rangle_{ij} \le 0, \qquad l_{ij}^- = \min\Big(1, \frac{Q_{ij}^-}{P_{ij}^-}\Big) \in [0,1], \tag{35a}$$

$$P_{ij}^+ = \sum_{(r,s)\in\mathcal{S}(i,j)} \max\left(A_{ij}^{rs}, 0\right) \ge 0, \qquad Q_{ij}^+ = M - \langle u_{LO}^{(n+1)}\rangle_{ij} \ge 0, \qquad l_{ij}^+ = \min\Big(1, \frac{Q_{ij}^+}{P_{ij}^+}\Big) \in [0,1], \tag{35b}$$

where m and M are the lower and upper bounds in (3) we want to impose to $\langle u_h^{(n+1)}\rangle_{ij}$. The new update of the mean value of the solution is now defined by:

$$\langle u_h^{(n+1)}\rangle_{ij} - \langle u_{h,LO}^{(n+1)}\rangle_{ij} = \sum_{(r,s)\in\mathcal{S}(i,j)} l_{ij}^{rs} A_{ij}^{rs}, \quad l_{ij}^{rs} = \begin{cases} \min(l_{ij}^-, l_{rs}^+) & \text{if } A_{ij}^{rs} < 0 \\ \min(l_{rs}^-, l_{ij}^+) & \text{otherwise.} \end{cases} \tag{36}$$

As a consequence, the cell-averaged solution satisfies the maximum principle (see [20, Lemma 5.4]): using (36) we get

$$\langle u_h^{(n+1)}\rangle_{ij} - \langle u_{h,LO}^{(n+1)}\rangle_{ij} \ge \sum_{(r,s)\in\mathcal{S}(i,j)} l_{ij}^{rs} \min(A_{ij}^{rs}, 0) = \min(l_{ij}^-, l_{rs}^+) P_{ij}^- \ge l_{ij}^- P_{ij}^- \ge Q_{ij}^- = m - \langle u_{LO}^{(n+1)}\rangle_{ij},$$
$$\langle u_h^{(n+1)}\rangle_{ij} - \langle u_{h,LO}^{(n+1)}\rangle_{ij} \le \sum_{(r,s)\in\mathcal{S}(i,j)} l_{ij}^{rs} \max(A_{ij}^{rs}, 0) = \min(l_{rs}^-, l_{ij}^+) P_{ij}^+ \le l_{ij}^+ P_{ij}^+ \le Q_{ij}^+ = M - \langle u_{LO}^{(n+1)}\rangle_{ij},$$

since $Q_{ij}^- \le l_{ij}^- P_{ij}^- \le 0$ and $0 \le l_{ij}^+ P_{ij}^+ \le Q_{ij}^+$ by definition (35).

Likewise, by (36) we have $l_{ij}^{rs} = l_{rs}^{ij}$ for $(r,s) \in \mathcal{S}(i,j)$, thus ensuring conservation of the method:

$$\sum_{ij} \langle u_h^{(n+1)}\rangle_{ij} = \sum_{ij} \langle u_{h,HO}^{(n+1)}\rangle_{ij} = \sum_{ij} \langle u_{h,LO}^{(n+1)}\rangle_{ij} = \sum_{ij} \langle u_h^{(n)}\rangle_{ij},$$

for periodic boundary conditions or compactly supported solutions.

From (36), the limiter is only applied at the interfaces and the DOFs can be evaluated explicitly from $u_{h,LO}^{(n+1)}$ and $u_{h,HO}^{(n+1)}$ through

$$\begin{aligned}\frac{\omega_k\omega_l}{4}(U_{ij}^{kl,n+1} - U_{ij,HO}^{kl,n+1}) = {} & \delta_{kp}\frac{\omega_l\lambda_{x_i}}{2}(1 - l_{ij}^{(i+1)j})\Big(U_{ij,HO}^{pl,n+1} - U_{ij,LO}^{pl,n+1}\Big) - \delta_{k0}\frac{\omega_l\lambda_{x_i}}{2}(1 - l_{ij}^{(i-1)j})\Big(U_{(i-1)j,HO}^{pl,n+1} - U_{(i-1)j,LO}^{pl,n+1}\Big) \\ & + \delta_{lp}\frac{\omega_k\lambda_{y_j}}{2}(1 - l_{ij}^{i(j+1)})\Big(U_{ij,HO}^{kp,n+1} - U_{ij,LO}^{kp,n+1}\Big) - \delta_{l0}\frac{\omega_k\lambda_{y_j}}{2}(1 - l_{ij}^{i(j-1)})\Big(U_{i(j-1),HO}^{kp,n+1} - U_{i(j-1),LO}^{kp,n+1}\Big).\end{aligned}$$

This limited scheme is conservative, satisfies the maximum principle for the cell-averaged solution, and is thus L^2-stable with the bound $\|u_h^{(n+1)}\|_{L^2(\Omega_h)} \le \sqrt{|\Omega_h|}\max(|m|,|M|)$, which provides existence of the solution, while uniqueness follows from uniqueness of the HO and LO solutions and the above explicit reconstruction for u_h^{n+1}. An entropy inequality seems however difficult to establish.

The FCT limiter requires to solve two linear systems for $u_{h,LO}^{(n+1)}$ and $u_{h,HO}^{(n+1)}$ at each time step. Let us stress that since we need to compute $u_{h,HO}^{(n+1)}$, we know easily if the limiter is required, that is if the maximum principle is violated for the cell-averaged solution in some cell of the mesh. If it is not violated, we set $u_h^{(n+1)} \equiv u_{h,HO}^{(n+1)}$ and do not need to compute $u_{h,LO}^{(n+1)}$, but only to apply the linear scaling limiter (see section 5.1). The FCT limiter may hence be viewed as an a posteriori limiter which is applied when needed after the solution update in the same way as other a posteriori limiters, such as in the MOOD method [9]. Preserving the maximum principle on the cell-averaged solution is a weaker requirement than preserving it on every DOFs and should therefore be more likely to be respected. As a consequence, the present FCT limiter is expected to less modify the solution, which is supported by the numerical experiments of section 5.3.

In the next section, we also propose efficient algorithms to solve these linear systems to mitigate the extra cost induced by the additional linear solution when $u_{h,LO}^{(n+1)}$ is required.

4.4. Linear system solution

Both linear systems without, (27), and with graph viscosity, (30), result in a block linear system

$$\mathbb{A}_{2d}\mathbf{U}^{(n+1)} = \mathbb{M}_{2d}\mathbf{U}^{(n)} \tag{37}$$

of size $N_xN_yN_p$ with blocks of size $N_p = (p+1)^2$ to be solved for $\mathbf{U}^{(n+1)}$ where $\mathbf{U}_{n_{kl}+(i-1)N_p+(j-1)N_xN_p} = U_{ij}^{kl}$ with $n_{kl} = 1 + k + l(p+1)$ and $\mathbb{M}_{2d}$ the global mass matrix. Considering the block structure of $\mathbb{A}_{2d}$ is important for efficiently solving (37) and usually requires the inversion of the diagonal blocks as a main step. These blocks are dense and hence require algorithms of complexity $O(N_p^3)$ for their inversion. We propose below algorithms based on the properties of the 1D schemes in section 3.4 for their efficient inversion. A repository of these algorithms (equations (40), (43), (45) and algorithm 1) is available at [1] and Appendix B provides a description of the repository together with some comparison of the performances of the different algorithms.

4.4.1. 1D diagonal blocks as building blocks of the 2D linear systems

Let us introduce the diagonalization in $\mathbb{C}$ of the matrix $\boldsymbol{\mathcal{L}}$ in (24):

$$\boldsymbol{\mathcal{L}} = \mathbf{R}\boldsymbol{\Psi}\mathbf{R}^{-1}, \tag{38}$$

where the columns of $\mathbf{R} \in \mathbb{C}^{(p+1)\times(p+1)}$ are the right eigenvectors of $\boldsymbol{\mathcal{L}}$ and $\boldsymbol{\Psi}$ is the diagonal matrix of the corresponding $p+1$ eigenvalues. We therefore have

$$\mathbf{L}_{1d} = \mathbf{R}\boldsymbol{\Psi}_\lambda\mathbf{R}^{-1}, \quad \boldsymbol{\Psi}_\lambda = \mathbf{I} - 2\lambda\boldsymbol{\Psi}, \tag{39}$$

for the 1D diagonal blocks in (24).

From (24), eigenpairs ψ and $\mathbf{r} = (r_0, \dots, r_p)^\top$, such that $\boldsymbol{\mathcal{L}}\mathbf{r} = \psi\mathbf{r}$, satisfy $\sum_l D_{lk}r_l - \delta_{kp}\frac{r_p}{\omega_p} = \psi r_k$ and summing this relation over $0 \le k \le p$ gives $-\frac{1}{\omega_p}r_p = \psi\sum_k r_k$ and for $\psi = 0$ we would have $r_p = 0$, hence $\mathbf{D}^\top\mathbf{r} = 0$ so $\mathbf{r} = 0$ since $\mathbf{D}^\top$ is of rank p. So we have $\psi \neq 0$ and we can invert the above relation with (18) to get $\mathbf{r} = -\frac{r_p}{\psi\omega_p}(\sum_{l=0}^p \psi^{-l}\mathbf{D}^l)^\top\mathbf{e}_p$ and the pth component with $r_p \neq 0$. ψ and $\mathbf{r}$ are thus given by

$$\omega_p\psi^{p+1} + \sum_{l=0}^{p}\psi^{p-l}D_{pp}^{(l)} = 0, \quad r_k = -\frac{1}{\omega_p}\sum_{l=0}^{p}\psi^{-l-1}D_{pk}^{(l)} \quad \forall 0 \le k \le p-1, \quad r_p = 1. \tag{40}$$

4.4.2. Diagonal blocks of the HO scheme (27)

Setting $\lambda = \lambda_{x_i} + \lambda_{y_j} > 0$, we rewrite the scheme (27) without graph viscosity as

$$\mathbf{L}_{2d}(\mathbf{M}\otimes\mathbf{M})\mathbf{U}_{ij}^{n+1} - \lambda_{x_i}(\mathbf{M}\otimes\mathbf{e}_0\mathbf{e}_p^\top)\mathbf{U}_{(i-1)j}^{n+1} - \lambda_{y_j}(\mathbf{e}_0\mathbf{e}_p^\top\otimes\mathbf{M})\mathbf{U}_{i(j-1)}^{n+1} = (\mathbf{M}\otimes\mathbf{M})\mathbf{U}_{ij}^{n}, \tag{41}$$

where the first matrix in the diagonal blocks may be written as follows from the definition of $\mathbf{L}_{1d}$ in (39):

$$\mathbf{L}_{2d} \coloneqq \frac{\lambda_{x_i}}{\lambda}\mathbf{I}\otimes\mathbf{L}_{1d} + \frac{\lambda_{y_j}}{\lambda}\mathbf{L}_{1d}\otimes\mathbf{I} = (\mathbf{R}\otimes\mathbf{R})\boldsymbol{\Psi}_{2d}(\mathbf{R}\otimes\mathbf{R})^{-1}, \quad \boldsymbol{\Psi}_{2d} = \frac{\lambda_{x_i}}{\lambda}\mathbf{I}\otimes\boldsymbol{\Psi}_\lambda + \frac{\lambda_{y_j}}{\lambda}\boldsymbol{\Psi}_\lambda\otimes\mathbf{I}, \tag{42}$$

with $\mathbf{I}$ the identity matrix in $\mathbb{R}^{p+1}$ and $\mathbf{L}_{1d}$ the 1D operator defined in (24). The diagonal matrix $\boldsymbol{\Psi}_{2d}$ has $1 - 2(\lambda_{x_i}\psi_l + \lambda_{y_j}\psi_k) \neq 0$ as n_{kl}th component. Hence the inverse of the diagonal blocks in (27) has an explicit expression

$$(\mathbf{M}\otimes\mathbf{M})^{-1}\mathbf{L}_{2d}^{-1} = \left((\mathbf{M}^{-1}\mathbf{R})\otimes(\mathbf{M}^{-1}\mathbf{R})\right)\boldsymbol{\Psi}_{2d}^{-1}(\mathbf{R}\otimes\mathbf{R})^{-1}, \tag{43}$$

$$\boldsymbol{\Psi}_{2d}^{-1} = \operatorname{diag}\left(\frac{1}{1-2(\lambda_{x_i}\psi_k + \lambda_{y_j}\psi_l)} : \ 1 \le n_{kl} = 1 + k + l(p+1) \le N_p\right).$$

Note that $\boldsymbol{\mathcal{L}}$ in (24) depends only on the approximation order of the scheme p, not on the λ_{x_i} and λ_{y_j}, so the matrices $\mathbf{R}$, $\mathbf{R}^{-1}$, $\mathbf{M}^{-1}\mathbf{R}$, $\boldsymbol{\Psi}$, $\boldsymbol{\Psi}_{2d}$, etc. may be computed once from (40) at the beginning of the computation.

Remark 4.2 (Fast Diagonalization Method). The above inversion strategy is not new. [33] proposed the so-called *Fast Diagonalization Method* (FDM) for solving generic PDEs discretized with tensor-product methods: (42) and (43) happen to be the specialization of such a framework to the DGSEM method at hand.

4.4.3. Diagonal blocks of the LO scheme (30)

Including the graph viscosity (31) into (27) modifies the diagonal blocks of the linear system and we now need to solve

$$\mathbf{L}^v_{2d}(\mathbf{M}\otimes\mathbf{M})\mathbf{U}^{n+1}_{ij} - \lambda_{x_i}(\mathbf{M}\otimes\mathbf{e}_0\mathbf{e}_p^\top)\mathbf{U}^{n+1}_{(i-1)j} - \lambda_{y_j}(\mathbf{e}_0\mathbf{e}_p^\top\otimes\mathbf{M})\mathbf{U}^{n+1}_{i(j-1)} = (\mathbf{M}\otimes\mathbf{M})\mathbf{U}^n_{ij}, \tag{44}$$

with

$$\mathbf{L}^v_{2d} = \mathbf{L}^0_{2d} - \mathbf{U}_v\mathbf{V}_v^\top, \tag{45}$$

and

$$\mathbf{L}^0_{2d} = \mathbf{L}_{2d} + 2d_{ij}\lambda\mathbf{I}\otimes\mathbf{I} = (\mathbf{R}\otimes\mathbf{R})(\mathbf{\Psi}_{2d} + 2d_{ij}\lambda\mathbf{I}\otimes\mathbf{I})(\mathbf{R}\otimes\mathbf{R})^{-1}, \quad \mathbf{U}_v = 2d_{ij}(\lambda_{x_i}\mathbf{I}\otimes\boldsymbol{\omega}, \lambda_{y_j}\boldsymbol{\omega}\otimes\mathbf{I}), \quad \mathbf{V}_v = (\mathbf{I}\otimes\mathbf{1}, \mathbf{1}\otimes\mathbf{I}), \tag{46}$$

where $\mathbf{U}_v$ and $\mathbf{V}_v$ are matrices in $\mathbb{R}^{N_p\times(2p+2)}$. Although the diagonal blocks $\mathbf{L}^v_{2d}$ may be efficiently built from the proposed method (45) (i.e., the 1D operators in $\mathbf{L}^0_{2d}$ plus a low-rank product) and then inverted with a direct solver, we propose below an alternative algorithm for their inversion that is found to be more efficient for polynomial degree up to $p \le 6$ (see Appendix B). Indeed, the matrix $\mathbf{L}^0_{2d}$ in (45) is easily inverted from (46) since $\mathbf{\Psi}_{2d} + 2d_{ij}\lambda\mathbf{I}\otimes\mathbf{I}$ is diagonal: here again, one takes advantage of the FDM. Then, we invert $\mathbf{L}^v_{2d}$ by using the Woodbury identity:

$$(\mathbf{L}^v_{2d})^{-1} \overset{(45)}{=} \Big(\mathbf{I}\otimes\mathbf{I} - (\mathbf{L}^0_{2d})^{-1}\mathbf{U}_v\mathbf{V}_v^\top\Big)^{-1}(\mathbf{L}^0_{2d})^{-1} = \Big(\mathbf{I}\otimes\mathbf{I} + (\mathbf{L}^0_{2d})^{-1}\mathbf{U}_v\Big(\mathbf{I}_{2p+2} - \mathbf{V}_v^\top(\mathbf{L}^0_{2d})^{-1}\mathbf{U}_v\Big)^{-1}\mathbf{V}_v^\top\Big)(\mathbf{L}^0_{2d})^{-1},$$

so the diagonal blocks may be inverted with algorithm 1, where only step 3 requires the inversion of a linear system of lower size with dense algebra tools. {Steps 1 and 2 can be solved with $O(4(p+1)^3)$ FLOPs: indeed, [33, Section 3] shows how to rewrite similar problems as a chain of cheap, low-size matrix-matrix multiplications (in step 1, a rewriting of $\mathbf{b}$ in matrix storage is needed) which is less expensive than solving the naive form. Finally, steps 3 and 4 require $O(4(p+1)N_p^2)$ and $O(2(p+1)N_p)$ FLOPs, respectively.

Algorithm 1 Algorithm flowchart for solving the system $\mathbf{L}^v_{2d}(\mathbf{M}\otimes\mathbf{M})\mathbf{x} = \mathbf{b}$ with graph viscosity.

1: solve $\mathbf{L}^0_{2d}\mathbf{y} = \mathbf{b}$ for $\mathbf{y}\in\mathbb{R}^{N_p}$ using (46):

$$\mathbf{y} = (\mathbf{R}\otimes\mathbf{R})\,\mathrm{diag}\left(\frac{1}{1+2\lambda d_{ij} - 2(\lambda_{x_i}\psi_k + \lambda_{y_j}\psi_l)} \,:\, n_{kl} \coloneqq 1 \le 1+k+l(p+1) \le N_p\right)(\mathbf{R}^{-1}\otimes\mathbf{R}^{-1})\mathbf{b};$$

2: solve $\mathbf{L}^0_{2d}\mathbf{Z} = \mathbf{U}_v$ for $\mathbf{Z}\in\mathbb{R}^{N_p\times(2p+2)}$ using (46):

$$\mathbf{Z} = 2d_{ij}(\mathbf{R}\otimes\mathbf{R})\,\mathrm{diag}\left(\frac{1}{1+2\lambda d_{ij} - 2(\lambda_{x_i}\psi_k + \lambda_{y_j}\psi_l)} \,:\, 1 \le n_{kl} \le N_p\right)(\lambda_{x_i}\mathbf{R}^{-1}\otimes(\mathbf{R}^{-1}\boldsymbol{\omega}), \lambda_{y_j}(\mathbf{R}^{-1}\boldsymbol{\omega})\otimes\mathbf{R}^{-1});$$

3: solve $(\mathbf{I}_{2p+2} - \mathbf{V}_v^\top\mathbf{Z})\mathbf{z} = \mathbf{V}_v^\top\mathbf{y}$ for $\mathbf{z}\in\mathbb{R}^{2p+2}$;
4: set $\mathbf{x} = (\mathbf{M}^{-1}\otimes\mathbf{M}^{-1})(\mathbf{y} + \mathbf{Z}\mathbf{z})$.

Remark 4.3 (Solution of the global system). In the case of Dirichlet boundary conditions, (37) is a block lower triangular system that can be efficiently solved with blockwise forward substitution by using the block inversion algorithms from the previous sections. In the case of periodic boundary conditions, it is possible to decompose the matrix into a block lower triangular matrix and a low-rank product: $\mathbb{A}_{2d} = \mathbb{A}^0_{2d} + \mathbf{U}_{2d}\mathbf{V}_{2d}^\top$ with

$$\mathbf{U}_{2d} = \frac{1}{2}\Big(\lambda_{y_1}(\omega_k\mathbf{e}^{k0}_{i1})^{0\le k\le p}_{1\le i\le N_x}, \lambda_{x_1}(\omega_l\mathbf{e}^{0l}_{1j})^{0\le l\le p}_{1\le j\le N_y}\Big), \quad \mathbf{V}_{2d} = \Big((\mathbf{e}^{k0}_{iN_y})^{0\le k\le p}_{1\le i\le N_x}, (\mathbf{e}^{0l}_{N_x j})^{0\le l\le p}_{1\le j\le N_y}\Big),$$

where the $\mathbf{e}^{kl}_{ij}$ are the components of the canonical basis of $\mathbb{R}^{N_xN_yN_p}$. It is therefore possible to apply the Woodbury identity in a similar way as in algorithm 1.

5. Numerical experiments

In this section we present numerical experiments on problems in one and two space dimensions (sections 5.2 and 5.3) in order to illustrate the properties of the DGSEM considered in this work. The FCT limiter (36) is applied in the 2D experiments only. A maximum principle holds for the cell-averaged solution, $m \leq \langle u_h^{(n+1)} \rangle \leq M$, in one space dimension and in two space dimensions with the FCT limiter. We then apply the linear scaling limiter from [58] described in section 5.1 to enforce a maximum principle on all the DOFs within the cells.

We evaluate error norms from the Gauss-Lobatto quadrature rules. Given $u_h \in \mathcal{V}_h^p$, we compute the L^2 norm as

$$\|u_h\|^2_{L^2(\Omega_h)} \coloneqq \sum_{i=1}^{N_x} \sum_{k=0}^{p} \frac{\Delta x}{2} \omega_k (U_i^k)^2, \quad \|u_h\|^2_{L^2(\Omega_h)} \coloneqq \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \sum_{k,l=0}^{p} \frac{\Delta x}{2} \frac{\Delta y}{2} \omega_k \omega_l (U_{ij}^{kl})^2,$$

in, respectively, 1D and 2D. The L^∞ norm is evaluated in a similar way:

$$\|u_h\|_{L^\infty(\Omega_h)} \coloneqq \max_{1\leq i\leq N_x} \max_{0\leq k\leq p} |U_i^k|, \quad \|u_h\|_{L^\infty(\Omega_h)} \coloneqq \max_{1\leq i\leq N_x} \max_{1\leq j\leq N_y} \max_{0\leq k,l\leq p} |U_{ij}^{kl}|.$$

5.1. Linear scaling limiter

Assuming $\langle u^{(n+1)} \rangle_\kappa \in [m, M]$ in a cell κ (either κ_i in 1D, or κ_{ij} in 2D), Zhang and Shu [58] proposed to modify $\mathbf{U}_\kappa^{(n+1)}$, the vector of DOFs in κ, as follows:

$$\widetilde{\mathbf{U}}_\kappa^{(n+1)} = \theta_\kappa \mathbf{U}_\kappa^{(n+1)} + (1-\theta_\kappa) \langle u_h^{(n+1)} \rangle_\kappa \mathbf{1}, \quad \theta_\kappa = \min\left(\left| \frac{M - \langle u_h^{(n+1)} \rangle_\kappa}{\max \mathbf{U}_\kappa - \langle u_h^{(n+1)} \rangle_\kappa} \right|, \left| \frac{m - \langle u_h^{(n+1)} \rangle_\kappa}{\min \mathbf{U}_\kappa - \langle u_h^{(n+1)} \rangle_\kappa} \right|, 1 \right), \tag{47}$$

with $\mathbf{1} = (1, 1, \ldots, 1)^\top \in \mathbb{R}^{(p+1)^d}$, and $\max \mathbf{U}_\kappa$ (resp., $\min \mathbf{U}_\kappa$) is the maximum (resp., minimum) value of the DOFs in the vector $\mathbf{U}_\kappa^{k,n+1}$. This limiter does not affect the high-order of accuracy for smooth solutions and does not change the cell average of the solution thus keeping the method conservative [58].

5.2. One space dimension

5.2.1. High-order accuracy in space

The high-order accuracy in space is first checked by looking for steady-state solutions of the following problem with a geometric source term and an inflow boundary condition:

$$\partial_t u + \partial_x u = 2\pi \cos(2\pi x) \quad \text{in } [0,1] \times [0,T], \quad u(0,\cdot) = 0 \quad \text{in } [0,T], \tag{48}$$

whose exact solution reads $u(x) = \sin(2\pi x)$. We take $\lambda = 1$, start from $u_0(x) = 0$, and march in time until $\|u_h^{n+1} - u_h^n\|_2 \leq 10^{-14}$. The $p+1$ accuracy of DGSEM is observed in Tab. 3. As expected [58, 60], the limiter does not affect the accuracy of the method.

5.2.2. Maximum-principle preservation

We now compare experiments with the theoretical bounds on the time to space steps ratio λ indicated in Tab. 1 for the DGSEM scheme to be maximum-principle preserving. We use a discontinuous initial condition composed of a Gaussian, a square pulse, a sharp triangle and a combination of semi-ellipses [31, Eq. (4.3)]. Table 4 displays the minimum and the maximum values of the cell average solution of (14) after a short physical time for different approximation orders and different values of λ. The results are in good agreement with the theoretical lower bounds in Tab. 1 and for $p \geq 2$ the maximum principle is seen to be violated on at least one mesh for the lowest value $\lambda = 0.1$.

In Tab. 5 we experimentally evaluate the lower bound on λ to guarantee a maximum principle on the cell-averaged solution by using a bisection method from the same configuration as in Tab. 4. We observe that the theoretical lower bound on λ derived in theorem 3.3 and Tab. 1 is sharp and confirmed by the experimental observations, though it seems to slightly overestimate the experimental lower bound for $p = 3$ and $p = 5$. Let us recall that the condition $\lambda > \lambda_{min}$ in Tab. 1 is a sufficient to obtain maximum principle preservation.

Table 3: Steady-state problem (48): $L^{k\in\{2,\infty\}}$ error levels $\|u_h - u\|_{L^k(\Omega_h)}$ and associated orders of convergence O_k obtained with $\lambda = 1$ when refining the mesh. The linear scaling limiter (47) is applied or not.

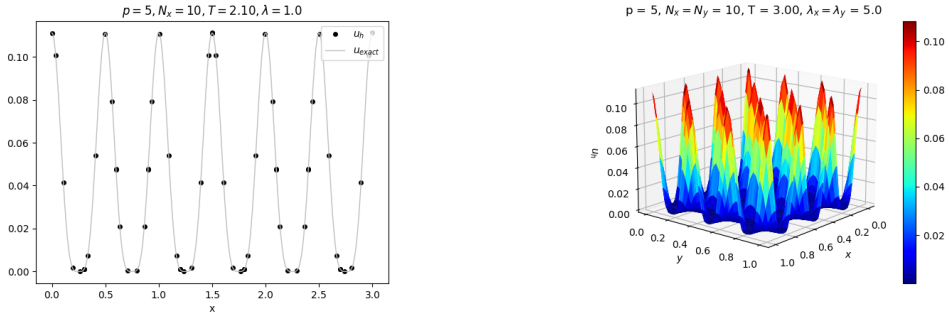
		no limiter				linear scaling limiter			
p	N_x	L^2 error	O_2	L^∞ error	O_∞	L^2 error	O_2	L^∞ error	O_∞
1	20	2.092E-2	-	4.071E-2	-	1.999E-2	-	4.071E-2	-
	40	5.239E-3	2.00	1.025E-2	1.99	5.120E-3	1.96	1.025E-2	1.99
	80	1.310E-3	2.00	2.569E-3	2.00	1.295E-3	1.98	2.569E-3	2.00
	160	3.276E-4	2.00	6.424E-4	2.00	3.258E-4	1.99	6.424E-4	2.00
2	20	4.164E-4	-	1.274E-3	-	4.309E-4	-	1.274E-3	-
	40	5.210E-5	3.00	1.609E-4	2.99	5.292E-5	3.03	1.609E-4	2.99
	80	6.515E-6	3.00	2.017E-5	3.00	6.564E-6	3.01	2.017E-5	3.00
	160	8.144E-7	3.00	2.523E-6	3.00	8.174E-7	3.01	2.523E-6	3.00
3	20	6.978E-6	-	2.669E-5	-	7.006E-6	-	2.669E-5	-
	40	4.365E-7	4.00	1.685E-6	3.99	4.367E-7	4.00	1.685E-6	3.99
	80	2.729E-8	4.00	1.056E-7	4.00	2.729E-8	4.00	1.056E-7	4.00
	160	1.706E-9	4.00	6.605E-9	4.00	1.706E-9	4.00	6.605E-9	4.00
4	20	1.008E-7	-	4.493E-7	-	1.008E-7	-	4.493E-7	-
	40	3.153E-9	5.00	1.418E-8	4.99	3.153E-9	5.00	1.418E-8	4.99
	80	9.854e-11	5.00	4.443e-10	5.00	9.854e-11	5.00	4.443e-10	5.00
	160	3.080e-12	5.00	1.390e-11	5.00	3.080e-12	5.00	1.390e-11	5.00
5	20	1.253E-9	-	6.274E-9	-	1.813E-9	-	1.249E-8	-
	40	1.959e-11	6.00	9.902e-11	5.99	2.446e-11	6.21	1.978e-10	5.98
	80	3.062e-13	6.00	1.550e-12	6.00	3.469e-13	6.14	3.103e-12	5.99
	160	4.504e-15	6.09	2.165e-14	6.16	4.757e-15	6.19	4.241e-14	6.19

Table 4: Linear scalar equation with a discontinuous initial condition: evaluation of the maximum principle for the cell-averaged solution as proved in theorem 3.3 and Tab. 1 after a short physical time $t = 0.01$. The solution should remain in the interval $[0, 1]$. The linear scaling limiter (47) is always applied.

		$N_x = 100$		$N_x = 101$	
p	λ	$\min_{1\le i\le N_x} \langle u_h\rangle_i$	$\max_{1\le i\le N_x} \langle u_h\rangle_i$	$\min_{1\le i\le N_x} \langle u_h\rangle_i$	$\max_{1\le i\le N_x} \langle u_h\rangle_i$
1	0.1	0.0	1.0	0.0	1.0
	0.25	9.88E-9	1.0	3.68E-10	1.0
	0.5	6.14E-6	1.0	9.40E-7	1.0
2	0.1	-3.92E-3	1.0005	-5.43E-3	1.005
	0.25	0.0	1.0	0.0	1.0
	0.5	5.35E-7	1.0	6.29E-8	1.0
3	0.1	0.0	1.0	-5.52E-3	1.006
	0.195137	0.0	1.0	0.0	1.0
	0.5	6.29E-7	1.0	8.91E-8	1.0
4	0.1	-4.00E-4	1.0002	-3.58E-5	1.000007
	0.151	0.0	1.0	0.0	1.0
	0.5	1.32E-7	1.0	8.93E-8	1.0
5	0.1	0.0	1.0	-2.45E-7	1.0005
	0.147568	0.0	1.0	0.0	1.0
	0.5	9.92E-8	1.0	9.06E-8	1.0
6	0.10	-1.78E-05	1.000013	-1.44E-4	1.000144
	0.109977	0.0	1.0	0.0	1.0
	0.5	0.0	1.0	0.0	1.0

Table 5: Experimental evaluation of the lower bounds of the time to space steps ratio λ_{min}^{exp} such that $\lambda_{1\leq i\leq N_x} = \lambda \geq \lambda_{min}^{exp}$ ensures the maximum principle preservation for the cell-averaged solution in theorem 3.3 and Tab. 1, while it doesn't for $\lambda \leq \lambda_{min}^{exp} - 10^{-2}$ on meshes with $N_x = 100$ and $N_x = 101$ elements. We report the theoretical values from Tab. 1 in the bottom line for the sake of comparison.

p	1	2	3	4	5	6
λ_{min}^{exp}	0	0.25	0.17	0.16	0.13	0.11
λ_{min}	0	0.25	0.195137	0.150346	0.147568	0.109977



(a) 1D problem

(b) 2D problem

Figure 2: Advection-reaction equation with source: (a) 1D steady-state DGSEM solution to (49) for $p = 5$, $N_x = 10$ and the linear scaling limiter (47); (b) 2D steady-state DGSEM solution for problem (50) with $p = 5$, $N_x = N_y = 10$ and the FCT limiter. The solution is plotted at quadrature points and T refers to the pseudo time required to converge the solution, i.e., $\|u_h^{n+1} - u_h^n\|_2 \leq 10^{-14}$.

5.2.3. Linear advection-reaction with source

We finally consider a linear advection-reaction problem with a geometric source term:

$$\partial_t u + c_x \partial_x u + \beta u = s(x) \quad \text{in } \Omega \times (0, \infty), \qquad u(x, 0) = u_0(x) \quad \text{in } \Omega. \tag{49}$$

with $\beta \geq 0$ and $s(\cdot) \geq 0$. Providing nonnegative initial and boundary data are imposed, the solution remains nonnegative for all time.

We here adapt the problem representative of the radiative transfer equations from [56, Ex. 6.2] with $c = 1$, $\beta = 6000$, $s(x) = \beta(\frac{1}{9}\cos^4(2\pi x) + \epsilon) - \frac{4}{9}\cos^3(2\pi x)\sin(2\pi x)$ on $\Omega = [0, 3]$, $\epsilon = 10^{-14}$, and an inflow boundary condition $u(0, t) = \frac{1}{9} + \epsilon$. This problem has a steady-state smooth solution, but with low positive values and large oscillations: $u(x) = \frac{1}{9}\cos^4(2\pi x) + \epsilon \geq \epsilon$ (see Fig. 2).

We again set $\lambda = 1$ and iterate up to convergence $\|u_h^{n+1} - u_h^n\|_2 \leq 10^{-14}$. Table 6 displays the error levels obtained for different approximation orders and mesh refinements when applying the scaling limiter (47) or not, together with the evaluation of the lowest value of the DGSEM solution. The limiter keeps the high-order accuracy of the DGSEM, while it successfully preserves positivity of the solution thus confirming that the DGSEM preserves positivity of the cell-averaged solution before the application of the limiter. We however observe a suboptimal pth order of accuracy with or without the limiter which was also reported in preceding experiments [8] and can be attributed to the low accuracy of the Gauss-Lobatto quadrature rules applied to the nonlinear geometric source term compared to other quadrature rules [10] (see also [8, Remark 3.1]).

5.3. Two space dimensions

We now focus on numerical tests in two space dimensions in the unit square using a Cartesian mesh with $N_x = N_y$ cells in the x and y directions respectively. For all the tests, we set $c_x = c_y = 1$. We here compare results obtained with the DGSEM scheme and without or with the FCT limiter:

no limiter: we solve (27) without graph viscosity for $u_h^{(n+1)}$. We cannot apply the linear scaling limiter (47) since the $\langle u_h^{(n+1)} \rangle_{ij}$ are not guaranteed to satisfy the maximum principle;

Table 6: Advection-reaction problem with source (49): $L^{k\in\{2,\infty\}}$ error levels $\|u_h - u\|_{L^k(\Omega_h)}$ and associated orders of convergence O_k obtained with $\lambda = 1$ when refining the mesh. The solution should remain in the interval $[0, \frac{1}{9}]$. Minimum value of DOFs $u_{h_{\min}} = \min(U^{0\le k\le p}_{1\le i\le N_x})$. The linear scaling limiter (47) is applied or not.

		no limiter					linear scaling limiter				
p	N_x	$u_{h_{\min}}$	L^2 error	O_2	L^∞ error	O_∞	$u_{h_{\min}}$	L^2 error	O_2	L^∞ error	O_∞
1	20	-5.28E-05	1.50E-04	–	1.71E-04	–	9.99E-15	1.49E-04	–	1.71E-04	–
	40	-1.04E-05	8.90E-05	0.76	1.02E-04	0.74	9.99E-15	8.42E-05	0.83	1.02E-04	0.74
	80	-1.46E-06	4.63E-05	0.94	5.35E-05	0.94	1.00E-14	4.43E-05	0.93	5.35E-05	0.94
	160	-3.01E-07	2.32E-05	1.00	2.70E-05	0.98	1.00E-14	2.26E-05	0.97	2.68E-05	1.00
2	20	-3.17E-05	5.52E-05	–	7.38E-05	–	1.00E-14	6.78E-05	–	1.57E-04	–
	40	-7.44E-06	1.58E-05	1.80	2.24E-05	1.72	1.00E-14	1.81E-05	1.91	4.35E-05	1.86
	80	-1.05E-06	4.11E-06	1.95	5.90E-06	1.93	1.00E-14	4.21E-06	2.10	6.51E-06	2.74
	160	-1.31E-07	1.03E-06	1.99	1.53E-06	1.94	1.00E-14	1.04E-06	2.02	1.53E-06	2.09
3	20	-5.33E-06	1.60E-05	–	2.71E-05	–	9.99E-15	1.71E-05	–	3.66E-05	–
	40	-1.46E-06	2.25E-06	2.83	3.84E-06	2.82	1.00E-14	2.71E-06	2.66	7.75E-06	2.24
	80	-2.59E-07	2.86E-07	2.98	4.81E-07	3.00	1.00E-14	3.30E-07	3.04	1.07E-06	2.85
	160	-3.36E-08	3.52E-08	3.02	6.22E-08	2.95	1.00E-14	3.81E-08	3.11	1.35E-07	2.99
4	20	-5.26E-06	3.72E-06	–	6.56E-06	–	1.00E-14	5.68E-06	–	2.05E-05	–
	40	-2.65E-07	2.58E-07	3.85	4.55E-07	3.85	1.00E-14	3.08E-07	4.21	1.33E-06	3.94
	80	-8.99E-09	1.66E-08	3.96	3.11E-08	3.87	1.00E-14	1.71E-08	4.17	4.76E-08	4.81
	160	-2.73E-10	1.04E-09	3.99	2.08E-09	3.90	1.00E-14	1.04E-09	4.03	2.08E-09	4.51
5	20	-3.30E-07	7.12E-07	–	1.32E-06	–	9.99E-15	7.78E-07	–	1.32E-06	–
	40	-2.62E-08	2.41E-08	4.88	4.58E-08	4.8	1.00E-14	3.05E-08	4.67	7.84E-08	4.08
	80	-1.06E-09	7.54E-10	5.00	1.38E-09	5.04	1.00E-14	9.22E-10	5.05	3.10E-09	4.66
	160	-3.14E-11	2.29E-11	5.04	4.63E-11	4.91	1.00E-14	2.56E-11	5.17	9.06E-11	5.10

Table 7: Verification of the maximum principle for problem (25) after one time step on a mesh with $N_x \times N_y = 20 \times 20$ elements, $\lambda_{x_i} = \lambda_{y_j} = \lambda$, and the discontinuous initial condition $u_0(x, y) = 1_{|x-\frac{1}{4}|+|y-\frac{1}{4}|\leq 0.15}$. The solution should remain in the interval $[0, 1]$.

		no limiter		FCT limiter	
p	λ	$\min_{1\leq i,j\leq 20} \langle u_h \rangle_{ij}$	$\max_{1\leq i,j\leq 20} \langle u_h \rangle_{ij}$	$\min_{1\leq i,j\leq 20} \langle u_h \rangle_{ij}$	$\max_{1\leq i,j\leq 20} \langle u_h \rangle_{ij}$
	0.05	0.0	1.0	0.0	1.0
1	1	-9.45E-3	0.90	9.59E-08	0.90
	5	-9.45E-3	0.47	1.37E-02	0.49
	0.05	-6.76E-4	1.0002	0.0	1.0
2	1	-6.76E-3	0.93	1.42E-07	0.90
	5	-6.60E-3	0.44	2.01E-3	0.41
	0.05	-4.85E-8	1.0	0.0	1.0
3	1	-2.98E-3	0.92	5.40E-07	0.91
	5	-6.24E-4	0.44	3.22E-03	0.38
	0.05	-1.41E-4	1.0007	0.0	1.0
4	1	-1.33E-4	0.92	5.97E-07	0.92
	5	-6.09E-5	0.44	3.25E-03	0.38
	0.05	-1.31E-3	1.0	0.0	1.0
5	1	-8.80E-5	0.92	6.45E-07	0.92
	5	-1.10E-4	0.44	3.33E-03	0.38

FCT limiter: we first solve (27) without graph viscosity for $u_{h,HO}^{(n+1)}$ and check if the cell-averaged solution satisfies the maximum principle, and if so, we set $u_h^{(n+1)} \equiv u_{h,HO}^{(n+1)}$. If not, we solve (30) with graph viscosity for $u_{h,LO}^{(n+1)}$ and apply the FCT limiter (36) introduced in section 4.3. Finally, we apply the linear scaling limiter (47) after the FCT limiter to preserve a maximum principle on all the DOFs in $u_h^{(n+1)}$.

Let us recall that the association of the FCT and linear scaling limiters has been introduced to avoid excessive limiting. We refer to Appendix A for a comparison with the FCT limiter only applied to all DOFs.

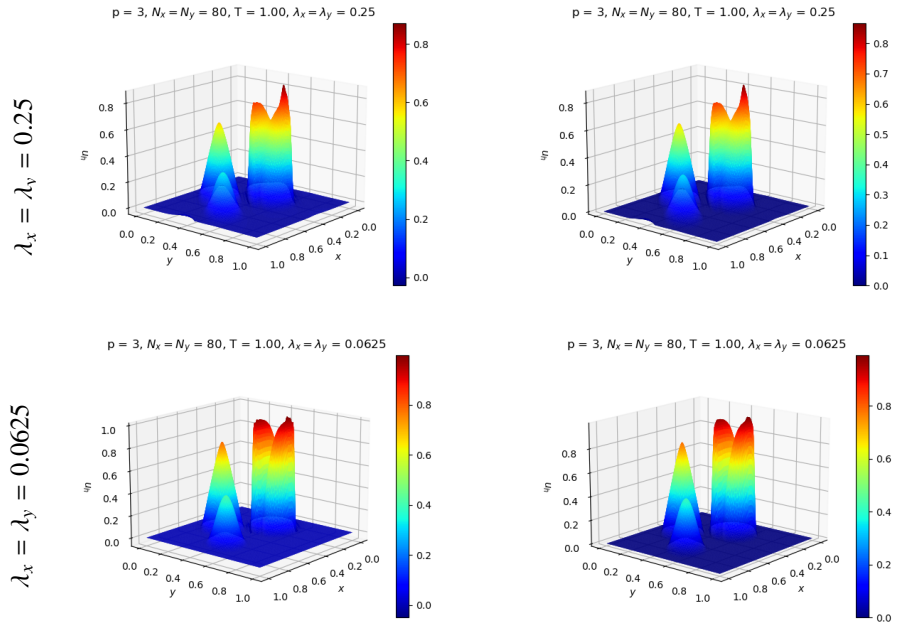
5.3.1. Maximum-principle preservation

We first evaluate both DGSEM schemes on an unsteady problem with a discontinuous initial condition, $u_0(x, y) = 1$ if $|x - \frac{1}{4}| + |y - \frac{1}{4}| \leq 0.15$ and 0 else, and periodic boundary conditions. Table 7 gives the minimum and maximum values of the cell-averaged solution after one time step with $1 \leq p \leq 5$ and different values of $\lambda_x = \lambda_y$. The maximum principle is not satisfied when using the DGSEM without limiter, except for $p = 1$ and the smallest time step. In particular, the maximum principle is violated even for large $\lambda_x = \lambda_y$ in contrast to what is observed and proved in one space dimension. As expected, the FCT limiter successfully imposes a maximum principle on the cell-averaged solution, thus enabling a maximum principle through the use of the linear scaling limiter.

We now consider the transport of a nonsmooth solution and solve (25) with periodic boundary conditions and the initial condition from the solid body rotation test case which contains a smooth bump, a cone and a slotted disk (see [57, 29, 28] for details). Fig. 3 displays fourth-order solutions obtained on a coarse mesh with and without the FCT limiter and for two different time steps. As expected, the temporal error induced by the backward Euler integration dominates and the solution with the low time step presents a better resolution of the different features. The FCT limiter successfully imposes the maximum principle without destroying the resolution capabilities of the scheme.

5.3.2. Steady smooth solution

We now consider a smooth steady-state solution of the problem $\partial_x u + \partial_y u = 0$ in $\Omega = [0, 1]^2$ with inlet conditions $u(x, 0) = \sin(2\pi x)$, $u(0, y) = -\sin(2\pi y)$ and outflow conditions at boundaries $x = 1$ and $y = 1$. The exact solution is $u(x, y) = \sin(2\pi(x - y))$. In practice, we look for a steady solution to the unsteady problem (1). We take $\lambda_x = \lambda_y = 5$, start from $u_h^{(0)} \equiv 0$ and march in time until $\|u_h^{n+1} - u_h^n\|_2 \leq 10^{-14}$ with the DGSEM scheme with FCT limiter. Error levels are summarized in Tab. 8 together with the minimum and maximum values of the cell-averaged solution. The



(a) no limiter (b) FCT limiter

Figure 3: Nonsmooth initial condition: DGSEM solutions at time $T = 1$ for a nonsmooth initial condition obtained with $N_x = N_y = 80$, $p = 3$, two different time steps, without (left), or with (right) the FCT limiter. The solution is plotted at quadrature points.

FCT limiter keeps here the $p + 1$ high-order accuracy in space of the DGSEM while it successfully preserves the maximum principle on the cell-averaged solution, and hence also on the DOFs through the linear scaling limiter.

5.3.3. Steady discontinuous solution

We now consider a discontinuous steady solution and consider $\partial_x u + \partial_y u = 0$ in $\Omega = [0,1]^2$, inlet conditions $u(x,0) = \cos(\pi x)$, $u(0,y) = -\cos(\pi y)$ and outflow conditions at boundaries $x = 1$ and $y = 1$. The exact solution is $u(x,y) = \mathrm{sgn}(x-y)\cos(\pi(x-y))$, with sgn the sign function, and is therefore discontinuous at $x = y$. Results are reported in Tab. 9 and Fig. 4. Here again, the FCT limiter is required to guarantee the maximum principle. In particular, the DGSEM without limiter violates the maximum principle for the cell-averaged solution which prevents the use of the linear scaling limiter.

5.3.4. Linear advection-reaction with source

We finally consider a linear advection-reaction problem with a geometric source term:

$$\partial_x u + \partial_y u + \beta u = s(x,y) \quad \text{in } \Omega, \qquad u(x,0) = u_0(x) \quad \text{in } \Omega. \tag{50}$$

with $\beta \geq 0$, $s(\cdot,\cdot) \geq 0$, and nonnegative inflow boundary data. We adapt the problem from section 5.2.3 and [56] to two space dimensions and set $\beta = 6000$ and a source term $s(x,y)$ such that the solution is $u(x,y) = \frac{1}{9}\cos(3\pi x)^4 cos(3\pi y)^4$ (see Fig. 2). Inflow boundary conditions, $u(x,0) = \frac{1}{9}\cos(3\pi x)^4$ and $u(0,y) = \frac{1}{9}\cos(3\pi y)^4$, are applied to $x = 0$ and $y = 0$, while outflow conditions are imposed at $x = 1$ and $y = 1$.

Tab. 10 displays the error levels together with minimum and maximum values of the DOFs obtained without or with the FCT limiter, different approximation orders and different mesh refinements. We again use $\lambda_x = \lambda_y = 5$, start from $u_h^{(0)} \equiv 0$ and march in time until $\|u_h^{n+1} - u_h^n\|_2 \leq 10^{-14}$. As in the 1D case in section 5.2.3, we observe a suboptimal convergence order of p as the mesh is refined due to the insufficient accuracy of the Gauss-Lobatto quadrature rules for integrating the highly nonlinear geometric source terms. Using the limiter or not leads to comparable error levels, while the FCT limiter is necessary for the solution to satisfy the maximum principle.

Table 8: Smooth steady-state problem: $L^{k\in\{2,\infty\}}$ error levels $\|u_h - u\|_{L^k(\Omega_h)}$ and associated orders of convergence O_k for problem $\partial_x u + \partial_y u = 0$ with data $u(x,0) = \sin(2\pi x)$, $u(0,y) = -\sin(2\pi y)$ obtained with $\lambda_x = \lambda_y = 5$ when refining the mesh and using the FCT limiter. The solution should remain in the interval $[-1,1]$. Minimum and maximum values of the cell-averaged solution over the mesh: $\langle u_h\rangle_{\min\,/\,\max} = \min\,/\,\max(\langle u_h\rangle_{ij} : 1 \le i \le N_x, 1 \le j \le N_y)$.

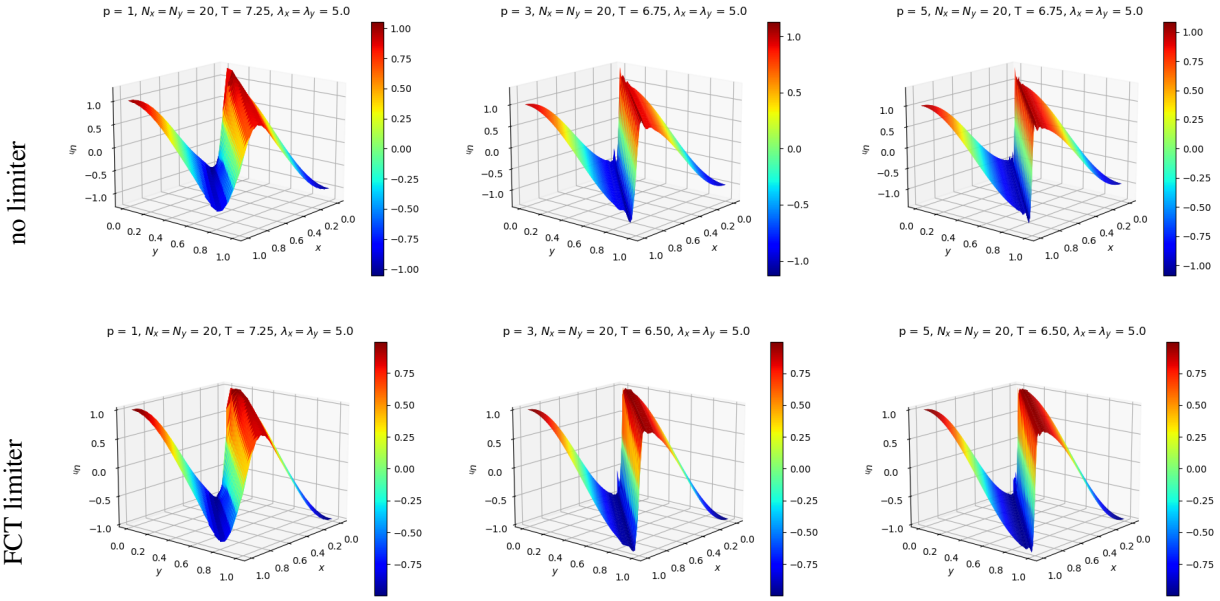
p	$N_x = N_y$	$\langle u_h\rangle_{\min}$	$\langle u_h\rangle_{\max}$	L^2 error	O_2	L^∞ error	O_∞
1	5	-0.7803	0.7803	$3.260E-01$	–	6.805E-01	–
	10	-0.9313	0.9313	9.840E-02	1.73	2.779E-01	1.29
	20	-0.9955	0.9955	2.431E-02	2.02	6.341E-02	2.13
	40	-0.9991	0.9991	6.589E-03	1.88	1.789E-02	1.83
2	5	-0.8293	0.8293	3.808E-02	–	1.610E-01	–
	10	-0.9200	0.9200	4.770E-03	3.00	1.348E-02	3.58
	20	-0.9917	0.9917	6.038E-04	2.98	2.354E-03	2.52
	40	-0.9979	0.9979	7.377E-05	3.03	2.084E-04	3.50
3	5	-0.8322	0.8322	2.511E-03	–	8.746E-03	–
	10	-0.9201	0.9201	1.569E-04	4.00	7.599E-04	3.52
	20	-0.9918	0.9918	1.074E-05	3.87	7.432E-05	3.35
	40	-0.9979	0.9979	6.457E-07	4.06	4.724E-06	3.98
4	5	-0.8323	0.8323	1.430E-04	–	6.283E-04	–
	10	-0.9201	0.9201	4.545E-06	4.98	1.880E-05	5.06
	20	-0.9918	0.9918	1.431E-07	4.99	6.162E-07	4.93
	40	-0.9979	0.9979	4.461E-09	5.00	1.950E-08	4.98
5	5	-0.8323	0.8323	7.131E-06	–	3.774E-05	–
	10	-0.9201	0.9201	1.131E-07	5.98	7.490E-07	5.65
	20	-0.9918	0.9918	4.074E-09	4.80	6.652E-08	3.49
	40	-0.9979	0.9979	4.789E-11	6.41	1.058E-09	5.97

Table 9: Discontinuous steady-state problem: verification of the maximum principle for problem $\partial_x u + \partial_y u = 0$ with data $u(x,0) = \cos(\pi x)$, $u(0,y) = -\cos(\pi y)$ obtained with $\lambda_x = \lambda_y = 5$ without and with the FCT limiter, and with $N_x = N_y = N$. The solution should remain in the interval $[-1,1]$. Minimum and maximum values of the cell-averaged solution and DOFs over the mesh: $\langle u_h\rangle_{\min\,/\,\max} = \min\,/\,\max(\langle u_h\rangle_{ij} : 1 \le i \le N_x, 1 \le j \le N_y)$ and $u_{h_{\min\,/\,\max}} = \min\,/\,\max(U^{0\le k,l\le p}_{1\le i,j\le N})$.

		no limiter				FCT limiter			
N	p	$\langle u_h\rangle_{\min}$	$\langle u_h\rangle_{\max}$	$u_{h_{\min}}$	$u_{h_{\max}}$	$\langle u_h\rangle_{\min}$	$\langle u_h\rangle_{\max}$	$u_{h_{\min}}$	$u_{h_{\max}}$
5	1	-0.7518	0.7518	-1.1363	1.1363	-0.7512	0.7512	-1.0000	1.0000
	2	-0.7820	0.7820	-1.2634	1.2634	-0.7820	0.7820	-1.0000	1.0000
	3	-0.7972	0.7972	-1.3364	1.3364	-0.7827	0.7827	-1.0000	1.0000
	4	-0.7832	0.7832	-1.3633	1.3633	-0.7828	0.7828	-1.0000	1.0000
	5	-0.7828	0.7828	-1.3764	1.3764	-0.7828	0.7828	-1.0000	1.0000
20	1	-1.0121	1.0121	-1.2437	1.2437	-0.9967	0.9967	-1.0000	1.0000
	2	-1.0465	1.0465	-1.2843	1.2843	-0.9781	0.9781	-1.0000	1.0000
	3	-1.0042	1.0042	-1.3438	1.3438	-0.9857	0.9857	-1.0000	1.0000
	4	-0.9937	0.9937	-1.3667	1.3667	-0.9857	0.9857	-1.0000	1.0000
	5	-0.9857	0.9857	-1.3781	1.3781	-0.9857	0.9857	-1.0000	1.0000

Table 10: 2D advection-reaction with source: $L^{k\in\{2,\infty\}}$ error levels $\|u_h - u\|_{L^k(\Omega_h)}$ and associated orders of convergence $\mathcal{O}_k$ for problem (50) with data $u(x,0) = \frac{1}{9}\cos(3\pi x)^4$, $u(0,y) = \frac{1}{9}\cos(3\pi y)^4$ obtained with $\lambda_x = \lambda_y = 5$ when refining the mesh with $N_x = N_y = N$. The solution should remain in the interval $[0, \frac{1}{9}]$. Minimum and maximum values of the DOFs: $u_{h_{\min\,/\,\max}} = \min\,/\,\max(U_{1\le i,j\le N}^{0\le k,l\le p})$.

		no limiter						with FCT limiter					
p	N	$u_{h_{\min}}$	$u_{h_{\max}}$	L^2 error	$\mathcal{O}_2$	L^∞ error	$\mathcal{O}_\infty$	$u_{h_{\min}}$	$u_{h_{\max}}$	L^2 error	$\mathcal{O}_2$	L^∞ error	$\mathcal{O}_\infty$
1	5	7.395e-06	0.1113	1.210E-04	–	2.953E-04	–	7.395E-06	0.1111	1.212E-04	–	2.953E-04	–
	10	-7.913e-05	0.1114	4.220E-04	−1.80	4.220E-04	−0.51	0.0000	0.1111	4.220E-04	−1.80	4.220E-04	−0.52
	20	-1.568e-05	0.1114	5.693E-05	2.89	2.933E-04	0.52	0.0000	0.1111	5.634E-05	2.91	2.933E-04	0.52
	40	-2.206e-06	0.1113	2.958E-05	0.95	1.581E-04	0.89	0.0000	0.1111	1.581E-04	0.94	1.581E-04	0.89
2	5	-9.775e-08	0.1116	8.671E-05	–	4.666E-04	–	0.0000	0.1111	8.539E-05	–	4.666E-04	–
	10	-4.737e-05	0.1113	3.532E-05	1.30	2.196E-04	1.09	0.0000	0.1111	4.664E-05	0.87	4.471E-04	0.06
	20	-1.107e-05	0.1112	1.014E-05	1.80	5.996E-05	1.87	0.0000	0.1111	1.132E-05	2.04	9.358E-05	2.26
	40	-1.557e-06	0.1111	2.630E-06	1.95	1.475E-05	2.02	0.0000	0.1111	2.671E-06	2.08	1.759E-05	2.41
3	5	2.192e-07	0.1114	5.032E-05	–	2.604E-04	–	2.192E-07	0.1111	5.215E-05	–	2.596E-04	–
	10	-8.469e-06	0.1111	1.029E-05	2.29	6.681E-05	1.96	0.0000	0.1111	1.217E-05	2.10	1.248E-04	1.06
	20	-2.173e-06	0.1111	1.440E-06	2.84	1.091E-05	2.61	0.0000	0.1111	1.722E-06	2.82	1.443E-05	1.11
	40	-3.801e-07	0.1111	1.817E-07	2.99	1.399E-06	2.96	0.0000	0.1111	2.023E-07	3.09	1.693E-06	3.09
4	5	-2.96521e-05	0.1112	2.500E-05	–	9.753E-05	–	0.0000	0.1111	3.727E-05	–	3.127E-04	–
	10	-7.80861e-06	0.1111	2.375E-06	3.40	1.906E-05	2.36	0.0000	0.1111	4.595E-06	3.02	6.962E-05	2.17
	20	-3.91095e-07	0.1111	1.648E-07	3.85	1.276E-06	3.90	0.0000	0.1111	1.999E-07	4.52	2.491E-06	4.80
	40	-1.30899e-08	0.1111	1.060E-08	3.96	8.080E-08	3.98	0.0000	0.1111	1.086E-08	4.20	1.009E-07	4.63
5	5	-1.13065e-06	0.1112	1.000E-05	–	7.942E-05	–	0.0000	0.1111	1.071E-05	–	9.570E-05	–
	10	-5.10247e-07	0.1111	4.557E-07	4.45	3.288E-06	4.59	0.0000	0.1111	5.795E-07	4.20	6.230E-06	3.94
	20	-3.81951e-08	0.1111	1.538E-08	4.89	1.301E-07	4.66	0.0000	0.1111	2.017E-08	4.85	1.740E-07	5.16
	40	-1.49843e-09	0.1111	4.767E-10	5.01	3.933E-09	5.05	0.0000	0.1111	5.647E-10	5.16	5.807E-09	4.91



(a) $p = 1$ (b) $p = 3$ (c) $p = 5$

Figure 4: Discontinuous steady-state problem: DGSEM solutions for a discontinuous steady-state problem $\partial_x u + \partial_y u = 0$ with data $u(x, 0) = \cos(\pi x)$, $u(0, y) = -\cos(\pi y)$ obtained with $\lambda_x = \lambda_y = 5$, $N_x = N_y = 20$, without and with the FCT limiter. The solution is plotted at quadrature points and T refers to the pseudo time required to converge the solution, i.e., $\|u_h^{n+1} - u_h^n\|_2 \leq 10^{-14}$.

6. Concluding remarks

This work proposes an analysis of the high-order DGSEM discretization with implicit backward Euler time stepping for the approximation of hyperbolic linear scalar conservation equations in multiple space dimensions. Two main aspects are considered here. We first investigate the maximum principle preservation of the scheme. For the 1D scheme, we prove that the DGSEM preserves the maximum principle of the cell-averaged solution providing that the CFL number is larger than a lower bound. This result allows to use linear scaling limiters [58, 60] to impose all the DOFs to satisfy the maximum principle. This property however does not hold in multiple space dimensions and we propose to use the FCT limiter [20, 5, 57] to enforce the maximum principle on the cell-averaged solution, thus avoiding excessive limiting. The FCT limiter combines the DGSEM scheme with a low-order maximum-principle preserving scheme derived by adding graph viscosity to the DGSEM scheme. The linear scaling limiter is then used to impose the maximum principle to all the DOFs. Numerical experiments in one and two space dimensions are provided to illustrate the conclusions of the present analyses. Then, we investigate the inversion of the linear systems resulting from the time implicit discretization at each time step. We prove that the diagonal blocks are invertible and provide efficient algorithms for their inversion. Future work will concern the extension of this analysis to nonlinear hyperbolic scalar equations [44] and systems of conservation laws on unstructured grids, and to high-order time implicit integration. Another direction of research may consist in using the fast inversion algorithms introduced in this work for solving preconditionning steps based on tensor product of 1D building blocks in block-preconditionned iterative solvers.

Appendix A. FCT limiter to impose a maximum principle to all DOFs

We here present numerical results obtained with a FCT limiter (see section 4.3) designed to impose a maximum principle to all DOFs, not only the cell average. We provide results in 1D for the sake of clarity, though this limiter is

not required as highlighted in section 3. Likewise, we design the limiter to impose only a lower bound on the solution, $U_i^{k,n+1} \geq \min u_0(\cdot)$, without loss of generality.

The HO scheme is (14) which, by removing $\sum_{l=0}^{p} \omega_l D_{lk} U_{i,HO}^{l,n+1} - \delta_{kp} U_{i,HO}^{p,n+1} + \delta_{k0} U_{i,HO}^{0,n+1} = 0$ and using $\omega_k D_{kk} = \frac{\delta_{kp} - \delta_{k0}}{2}$, we rewrite under the skew-symmetric form [21] as

$$\frac{\omega_k}{2} U_{i,HO}^{k,n+1} + \lambda_i \Big(- \sum_{l=0}^{p} \omega_l (D_{lk} - \delta_{kl} D_{kk}) (U_{i,HO}^{l,n+1} + U_{i,HO}^{k,n+1}) + \delta_{kp} U_{i,HO}^{p,n+1} - \delta_{k0} U_{i-1,HO}^{p,n+1}) \Big) = \frac{\omega_k}{2} U_i^{k,n}, \tag{A.1}$$

We now compare two LO schemes:

LO_GV: the HO scheme with graph viscosity introduced in section 4.2 where $d_i = 2 \max_{0 \leq k \neq m \leq p} (-\frac{D_{mk}}{\omega_k})$ from (33):

$$\frac{\omega_k}{2} U_{i,LO}^{k,n+1} + \lambda_i \Big(- \sum_{l=0}^{p} \omega_l D_{lk} U_{i,LO}^{l,n+1} + \delta_{kp} U_{i,LO}^{p,n+1} - \delta_{k0} U_{i-1,LO}^{p,n+1} + d_i \omega_k \sum_{l=0}^{p} \omega_l (U_{i,LO}^{k,n+1} - U_{i,LO}^{l,n+1}) \Big) = \frac{\omega_k}{2} U_i^{k,n}, \tag{A.2}$$

LO_GV_SD: the HO scheme with graph viscosity and the sparse discrete derivative matrix $\tilde{\mathbf{D}}$, from [34] and defined in (34) (see remark 4.1), and $\omega_k \omega_l \tilde{d}_i = \frac{1}{2}$ satisfying (33):

$$\frac{\omega_k}{2} U_{i,LO}^{k,n+1} + \lambda_i \Big(- \sum_{l=\max(k-1,0)}^{\min(k+1,0)} \omega_l \tilde{D}_{lk} U_{i,LO}^{l,n+1} + \delta_{kp} U_{i,LO}^{p,n+1} - \delta_{k0} U_{i-1,LO}^{p,n+1} + \frac{1}{2} \sum_{l=\max(k-1,0)}^{\min(k+1,0)} (U_{i,LO}^{k,n+1} - U_{i,LO}^{l,n+1}) \Big) = \frac{\omega_k}{2} U_i^{k,n}, \tag{A.3}$$

and after easy manipulations, we rewrite (A.3) as a subcell finite volume scheme (using $U_{i,LO}^{-1,n+1} = U_{i-1,LO}^{p,n+1}$):

$$\frac{\omega_k}{2} U_{i,LO}^{k,n+1} + \lambda_i (U_{i,LO}^{k,n+1} - U_{i,LO}^{k-1,n+1}) = \frac{\omega_k}{2} U_i^{k,n}.$$

We rewrite the differences between the HO and LO schemes, then the limited and LO schemes as

$$\frac{\omega_k}{2} (U_{i,HO}^{k,n+1} - U_{i,LO}^{k,n+1}) = \sum_{(j,l) \in \mathcal{S}(i,k)} A_{ik}^{jl}, \qquad \frac{\omega_k}{2} (U_i^{k,n+1} - U_{i,LO}^{k,n+1}) = \sum_{(j,l) \in \mathcal{S}(i,k)} l_{ik}^{jl} A_{ik}^{jl},$$

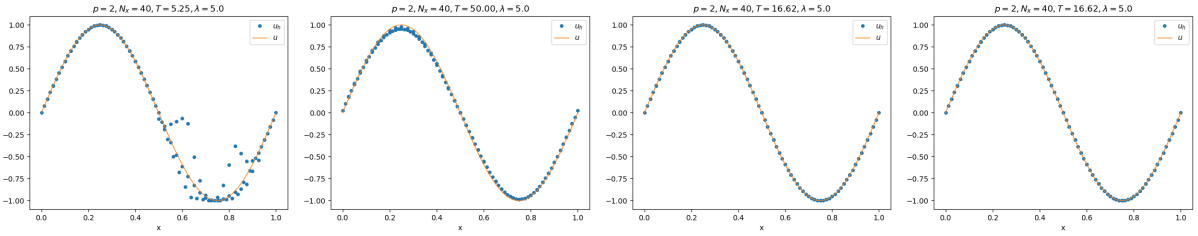
with $A_{ik}^{jl} = -A_{jl}^{ik}$ and where the limiter coefficients $l_{ik}^{jl} = l_{jl}^{ik}$ are computed from (35) and (36), but where $P_i^k = \sum_{(j,l)} \min(A_{ik}^{jl}, 0)$ and $Q_i^k = \frac{\omega_k}{2} (m - U_{i,LO}^{k,n+1})$.

Both limiters are now compared to the limiter introduced in section 4.3 (present limiter). Recall that the present limiter uses the same LO scheme as the LO_GV limiter, but applies the limiter to impose positivity of the cell-averaged solution, then applies the linear scaling limiter (47). For the sake of comparison, we also present results using (A.3) as LO scheme, but imposing positivity of the cell-averaged solution, then applying the linear scaling limiter (47). This last limiter will be referred as to **LO_GV_SD_aver**. For all limiters, we check if the minimum principle is satisfied by $u_{h,HO}^{(n+1)}$ and, if so, we set $u_h^{(n+1)} = u_{h,HO}^{(n+1)}$.

Figs. A.5 and A.6 compare the results for a smooth steady solution, while Fig. A.7 compares results for an unsteady discontinuous solution. Limiting all the DOFs directly with the FCT limiter affects the accuracy of the scheme by imposing an excessive limiting. This excessive limiting has already been reported [21, 23, 6]. It is worth noting that the limiter coefficients l_{ik}^{jl} in (35) scale with $\frac{1}{\lambda_i}$ and are thus expected to overlimit the solution for large time steps as observed in Fig. A.5. It is also observed that the limiter may also prevent convergence of the computation to steady-state as observed in Fig. A.6. Limiting the cell average does not suffer from these shortcomings and the results are less sensitive to the choice of the LO scheme.

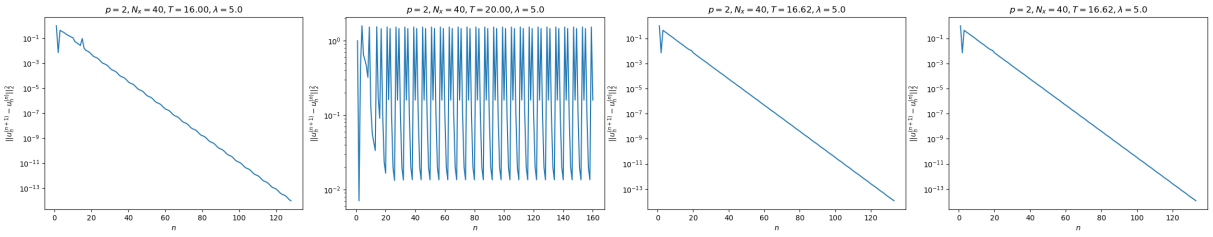
We end this section with a comparison of the accuracy of the LO schemes introduced above on the smooth test case. We also indicate the results obtained with a first-order finite volume scheme, that is with one DOF per cell. All LO schemes achieve first-order accuracy as expected, but the DGSEM scheme with graph viscosity and sparse discretization (A.3), which here reduces to a subcell finite volume scheme, provides quite lower error levels. This result is in agreement with those obtained with an explicit time stepping in [34, Fig. 1].

The main conclusions of the present tests are: (i) Imposing positivity of all DOFs directly with the FCT limiter successfully imposes positivity, but may affect the resolution capabilities of the scheme due to an excessive limiting. Especially with large time steps; (ii) Limiting the solution with the FCT limiter to impose positivity of the cell average only avoids these issues and make the results less sensitive to the choice of the LO scheme.



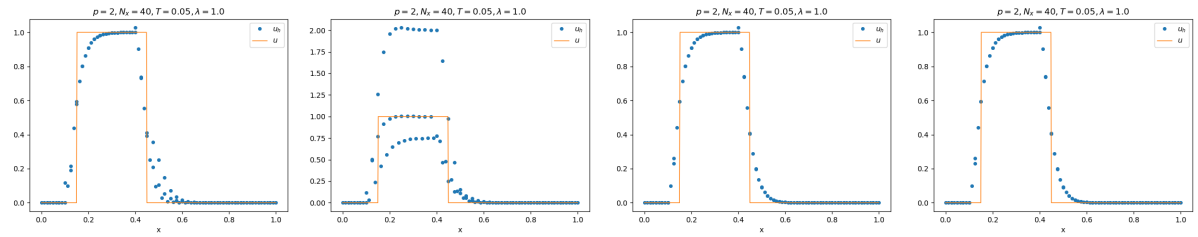
(a) LO_GV (b) LO_GV_SD (c) present limiter (d) LO_GV_SD_aver

Figure A.5: DGSEM approximation of a steady-state problem, $\partial_x u = \sin(2\pi x)$ and $u(0) = u(1)$, obtained with different FCT limiters ($p = 2$, $N_x = 40$, $\lambda_i = 5$). The solution is plotted at quadrature points and T refers to the pseudo time required to converge the solution, i.e., $\|u_h^{n+1} - u_h^n\|_2 \leq 10^{-14}$.



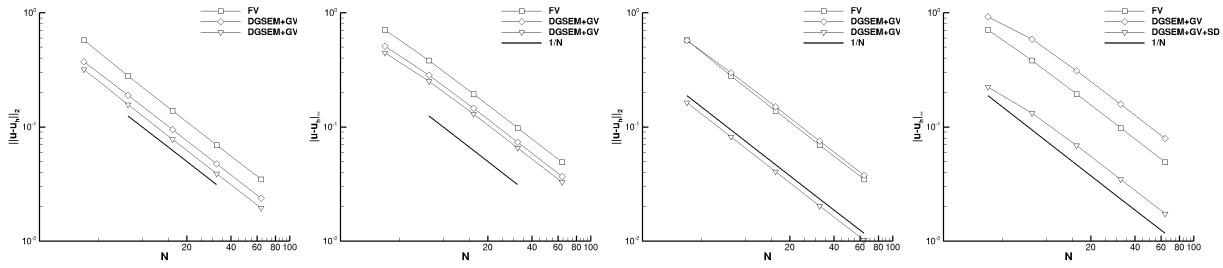
(a) LO_GV (b) LO_GV_SD (c) present limiter (d) LO_GV_SD_aver

Figure A.6: DGSEM approximation of a steady-state problem, $\partial_x u = \sin(2\pi x)$ and $u(0) = u(1)$, obtained with different FCT limiters ($p = 2$, $N_x = 40$, $\lambda_i = 5$). Residual histories $\|u_h^{n+1} - u_h^n\|_2$ as a function of time.



(a) LO_GV (b) LO_GV_SD (c) present limiter (d) LO_GV_SD_aver

Figure A.7: DGSEM approximation to an unsteady problem, $\partial_t u + \partial_x u = 0$, $u(0,t) = u(1,t)$ and $u_0(x) = 1_{[0.1,0.4]}$, obtained with different FCT limiters ($p = 2$, $N_x = 40$, $\lambda_i = 1$). The solution is plotted at quadrature points after two time steps.



(a) $p = 2$, L^2 error (b) $p = 2$, L^∞ error (c) $p = 4$, L^2 error (d) $p = 4$, L^∞ error

Figure A.8: Approximations of a steady-state problem, $\partial_x u = \sin(2\pi x)$ and $u(0) = u(1)$, obtained with different LO schemes: first-order finite volume scheme (FV); DGSEM scheme with graph viscosity (A.2) (DGSEM+GV); DGSEM scheme with graph viscosity and sparse discrete derivative matrix (A.3) (DGSEM+GV+SD). L^2 and L^∞ norms of the error obtained with $p = 2$ and $p = 4$.

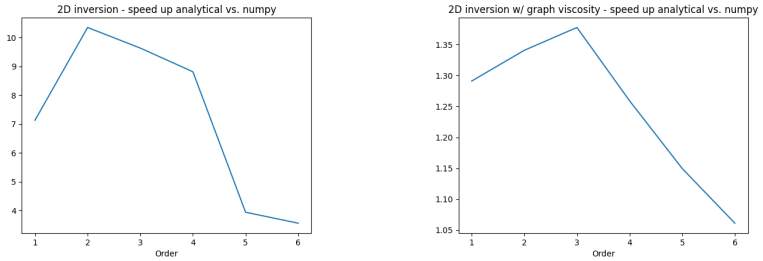


Figure B.9: Left: performance speed-up over polynomial degree for solving (B.1) using (43) with respect to standard algebraic tools (`numpy`). Right: similar to above, but for (B.2) and algorithm 1. Performance analysis has been evaluated thanks to built-in `python` module `timeit`; statistical data has been computed over 20 runs of 1000 resolution call each.

Appendix B. Inversion of diagonal blocks

The linear systems associated to the DGSEM discretization of problem (1) with an implicit time stepping have a sparse pattern with dense diagonal blocks of large size. Block-based iterative and exact solvers require the inversion of the diagonal blocks and we propose efficient algorithms to speed up the inversion of the diagonal blocks with respect to standard inversion algorithms. We implemented the proposed methods and compared them with standard ones. The code is freely available online [1]. It is written in `python` by using linear algebra tools of the `numpy` library [24]. We here focus on the linear system obtained when considering the 2D DGSEM problem which takes the form (see (41))

$$\mathbf{L}_{2d}(\mathbf{M} \otimes \mathbf{M})\mathbf{x} = \mathbf{b}, \tag{B.1}$$

where the involved matrices are defined in (42), and its counterpart obtained by adding the graph viscosity (see (44))

$$\mathbf{L}^{v}_{2d}(\mathbf{M} \otimes \mathbf{M})\mathbf{x} = \mathbf{b} \tag{B.2}$$

see (45) and (46) for the definition of the terms involved. The main goal of repository [1] is to assess the novel analytical way (43) (resp., algorithm 1) to solve (B.1) (resp., (B.2), by comparing it with the usage of reference algebraic tools (mainly, `numpy.linalg.inv`). We give in Fig. B.9 the performance comparisons obtained on a personal machine with 8 Intel Xeon(R) W-2223 CPUs and 16Gb RAM. One can reliably say that the novel inversion strategies (43) and algorithm 1 show consistent and often significant performance gains with respect to their dense counterparts. The gains are however less noticeable for high orders when solving (B.2).

Appendix C. Solution of the global 1D linear system

We here describe a fast algorithm to solve the global linear system (23). This is based on fast inversion of the diagonal blocks (24) with the Sherman-Morisson formula which provides the inverse of the sum of an invertible

matrix $\mathbf{A}$ and a rank-one matrix $\mathbf{u}\mathbf{v}^\top$: $(\mathbf{A}+\mathbf{u}\mathbf{v}^\top)^{-1} = (\mathbf{I} - \frac{1}{1+\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u}}\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^\top)\mathbf{A}^{-1}$. Using $\mathbf{A} = \mathbf{I} - 2\lambda\mathbf{D}^\top$ and $\mathbf{u} = \mathbf{v} = \mathbf{e}_p$, we obtain

$$\mathbf{M}^{-1}\mathbf{L}_{1d}^{-1} = \mathbf{M}^{-1}\left(\mathbf{I} - 2\lambda\left(\mathbf{D}^\top - \frac{1}{\omega_p}\mathbf{e}_p\mathbf{e}_p^\top\right)\right)^{-1} = \mathbf{M}^{-1}\left(\mathbf{I} - \frac{2\lambda}{\omega_p + 2\lambda\mathcal{D}_{pp}^i}\mathcal{D}\mathbf{e}_p\mathbf{e}_p^\top\right)\mathcal{D}$$

with $\mathcal{D} = (\mathbf{I} - 2\lambda\mathbf{D}^\top)^{-1}$ given by (18). This formula is well defined since $\omega_p + 2\lambda\mathcal{D}_{pp}^i > 0$ by (22). Let us propose a method to solve the global linear system (23). From (14), we observe that $\mathbb{A}_{1d} = \mathbb{A}_0 - \lambda_1\mathbf{e}_1^0(\mathbf{e}_{N_x}^p)^\top$ in (23) with $\mathbb{A}_0$ a block lower triangular matrix, with diagonal blocks $\mathbf{M}\mathbf{L}_{1d}$ and subdiagonal blocks $-\lambda_i\mathbf{e}_0\mathbf{e}_p^\top$, and a rank-one matrix defined from $(\mathbf{e}_i^k)_{1\le i\le N_x}^{0\le k\le p}$ the canonical basis of $\mathbb{R}^{N_x(p+1)}$. Using again the Sherman-Morisson formula, we easily solve (23) from algorithm 2 where steps 1 and 2 can be solved efficiently using blockwise forward substitution.

Algorithm 2 Algorithm flowchart for solving the global system (23) by using the decomposition $\mathbb{A}_{1d} = \mathbb{A}_0 - \lambda_1\mathbf{e}_1^0(\mathbf{e}_{N_x}^p)^\top$ with $\mathbb{A}_0$ a block lower triangular matrix and $\mathbf{e}_1^0(\mathbf{e}_{N_x}^p)^\top$ a rank-one matrix.

1: solve $\mathbb{A}_0\mathbf{V} = \mathbb{M}_{1d}\mathbf{U}^{(n)}$ for $\mathbf{V} \in \mathbb{R}^{N_x(p+1)}$;
2: solve $\mathbb{A}_0\mathbf{W} = \mathbf{e}_1^0$ for $\mathbf{W} \in \mathbb{R}^{N_x(p+1)}$;
3: set $\mathbf{U}^{(n+1)} = \mathbf{V} + \frac{\lambda_1\mathbf{e}_{N_x}^p\cdot\mathbf{V}}{1-\lambda_1\mathbf{e}_{N_x}^p\cdot\mathbf{W}}\mathbf{W}$.

Applying algorithm 2 again requires $1 - \lambda_1\mathbf{e}_{N_x}^p\cdot\mathbf{W} = 1 - \lambda_1\mathbf{e}_{N_x}^p\cdot(\mathbb{A}_0^{-1}\mathbf{e}_1^0) \neq 0$. This is indeed the case and to prove it we temporarily consider a uniform mesh for the sake of clarity, so $\lambda_i = \lambda$. We observe that the solution to $\mathbb{A}_0\mathbf{W} = \mathbf{e}_1^0$ satisfies $\mathbf{L}_{1d}\mathbf{M}\mathbf{W}_1 = \mathbf{e}_0$ and $\mathbf{L}_{1d}\mathbf{M}\mathbf{W}_i = (\lambda\mathbf{e}_0\mathbf{e}_p^\top)\mathbf{W}_{i-1}$ for $i \ge 2$. We thus get $\mathbf{W}_i = (\lambda\mathbf{M}^{-1}\mathbf{L}_{1d}^{-1}\mathbf{e}_0\mathbf{e}_p^\top)^{i-1}\mathbf{M}^{-1}\mathbf{L}_{1d}^{-1}\mathbf{e}_0$ and $\mathbf{e}_{N_x}^p\cdot\mathbf{W} = \lambda^{N_x-1}(\frac{1}{\omega_p}(\mathbf{L}_{1d}^{-1})_{p0})^{N_x}$ with $(\mathbf{L}_{1d}^{-1})_{p0} = \frac{2}{\omega_p}(1 - \frac{2\lambda}{\omega_p+2\lambda\mathcal{D}_{pp}^i}\mathcal{D}_{pp}^i)\mathcal{D}_{p0}^i = \frac{2\lambda\mathcal{D}_{p0}^i}{\omega_p+2\lambda\mathcal{D}_{pp}^i} > 0$ from (22). Note that $\mathcal{D}_{p0}^i > 0$ holds for $\lambda > \lambda_{min}$ defined in lemma 3.2. This latter condition is sufficient to apply algorithm 2, but is not necessary to invert the linear system (23) which is possible for all positive time steps.

References

[1] https://github.com/rueljean/fast_DGSEM_block_inversion.

[2] R. Abgrall, P. Rai, and F. Renac, *A discontinuous Galerkin spectral element method for a nonconservative compressible multicomponent flow model*, J. Comput. Phys., 472 (2023), p. 111693, https://doi.org/https://doi.org/10.1016/j.jcp.2022.111693.

[3] H. L. Atkins and C.-W. Shu, *Quadrature-free implementation of discontinuous Galerkin method for hyperbolic equations*, AIAA J., 36 (1998), pp. 775–782, https://doi.org/10.2514/2.436, https://doi.org/10.2514/2.436.

[4] M. Bohm, A. Winters, G. Gassner, D. Derigs, F. Hindenlang, and J. Saur, *An entropy stable nodal discontinuous Galerkin method for the resistive MHD equations. part i: Theory and numerical verification*, J. Comput. Phys., (2018).

[5] J. P. Boris and D. L. Book, *Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works*, J. Comput. Phys., 11 (1973), pp. 38–69, https://doi.org/https://doi.org/10.1016/0021-9991(73)90147-2.

[6] V. Carlier and F. Renac, *Invariant domain preserving high-order spectral discontinuous approximations of hyperbolic systems*, SIAM J. Sci. Comput., 45 (2023), pp. A1385–A1412, https://doi.org/10.1137/22M1492015.

[7] M. H. Carpenter, T. C. Fisher, E. J. Nielsen, and S. H. Frankel, *Entropy stable spectral collocation schemes for the Navier–Stokes equations: Discontinuous interfaces*, SIAM J. Sci. Comput., 36 (2014), pp. B835–B867.

[8] T. Chen and C.-W. Shu, *Entropy stable high order discontinuous Galerkin methods with suitable quadrature rules for hyperbolic conservation laws*, J. Comput. Phys., 345 (2017), pp. 427–461.

[9] S. Clain, S. Diot, and R. Loubère, *A high-order finite volume method for systems of conservation laws – multi-dimensional optimal order detection (MOOD)*, J. Comput. Phys., 230 (2011), pp. 4028–4050, https://doi.org/https://doi.org/10.1016/j.jcp.2011.02.026.

[10] B. Cockburn, S. Hou, and C. Shu, *The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. IV: The multidimensional case*, Math. Comp., 54 (1990), pp. 545–581, https://doi.org/https://doi.org/10.1090/S0025-5718-1990-1010597-0.

[11] F. Coquel, C. Marmignon, P. Rai, and F. Renac, *An entropy stable high-order discontinuous Galerkin spectral element method for the Baer-Nunziato two-phase flow model*, J. Comput. Phys., (2021), p. 110135, https://doi.org/https://doi.org/10.1016/j.jcp.2021.110135.

[12] A. Crivellini and F. Bassi, *An implicit matrix-free discontinuous galerkin solver for viscous and turbulent aerodynamic simulations*, Comput. Fluids, 50 (2011), pp. 81–93, https://doi.org/https://doi.org/10.1016/j.compfluid.2011.06.020.

[13] B. Després, *Entropy inequality for high order discontinuous Galerkin approximation of Euler equations*, in Hyperbolic Problems: Theory, Numerics, Applications, M. Fey and R. Jeltsch, eds., Basel, 1999, Birkhäuser Basel, pp. 225–231.

[14] L. T. Diosady and S. M. Murman, *Scalable tensor-product preconditioners for high-order finite-element methods: Scalar equations*, J. Comput. Phys., 394 (2019), pp. 759–776, https://doi.org/https://doi.org/10.1016/j.jcp.2019.04.047.

[15] A. Ern and J.-L. Guermond, *Invariant-domain preserving high-order time stepping: Ii. imex schemes*, SIAM J. Sci. Comput., 45 (2023), pp. A2511–A2538, https://doi.org/10.1137/22M1505025.

[16] T. C. Fisher and M. H. Carpenter, *High-order entropy stable finite difference schemes for nonlinear conservation laws: Finite domains*, J. Comput. Phys., 252 (2013), pp. 518–557.

[17] G. Gassner and D. A. Kopriva, *A comparison of the dispersion and dissipation errors of Gauss and Gauss-Lobatto discontinuous Galerkin spectral element methods*, SIAM J. Sci. Comput., 33 (2011), pp. 2560–2579, https://doi.org/10.1137/100807211.

[18] G. J. Gassner, *A skew-symmetric discontinuous Galerkin spectral element discretization and its relation to SBP-SAT finite difference methods*, SIAM J. Sci. Comput., 35 (2013), pp. A1233–A1253, https://doi.org/10.1137/120890144.

[19] G. J. Gassner, A. R. Winters, and D. A. Kopriva, *Split form nodal discontinuous Galerkin schemes with summation-by-parts property for the compressible Euler equations*, J. Comput. Phys., 327 (2016), pp. 39–66.

[20] J.-L. Guermond, M. Maier, B. Popov, and I. Tomas, *Second-order invariant domain preserving approximation of the compressible Navier–Stokes equations*, Comput. Methods Appl. Mech. Engrg., 375 (2021), p. 113608, https://doi.org/10.1016/j.cma.2020.113608.

[21] J.-L. Guermond, M. Nazarov, B. Popov, and I. Tomas, *Second-order invariant domain preserving approximation of the Euler equations using convex limiting*, SIAM J. Sci. Comput., 40 (2018), pp. A3211–A3239, https://doi.org/10.1137/17M1149961.

[22] J.-L. Guermond and B. Popov, *Invariant domains and first-order continuous finite element approximation for hyperbolic systems*, SIAM J. Numer. Anal., 54 (2016), pp. 2466–2489, https://doi.org/10.1137/16M1074291.

[23] J.-L. Guermond, B. Popov, and I. Tomas, *Invariant domain preserving discretization-independent schemes and convex limiting for hyperbolic systems*, Comput. Methods Appl. Mech. Engrg., 347 (2019), pp. 143–175, https://doi.org/https://doi.org/10.1016/j.cma.2018.11.036.

[24] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, *Array programming with NumPy*, Nature, 585 (2020), pp. 357–362, https://doi.org/10.1038/s41586-020-2649-2, https://doi.org/10.1038/s41586-020-2649-2.

[25] D. A. Kopriva, *Implementing Spectral Methods for Partial Differential Equations: Algorithms for Scientists and Engineers*, Springer Dordrecht, 2009, https://doi.org/https://doi.org/10.1007/978-90-481-2261-5.

[26] D. A. Kopriva and G. Gassner, *On the quadrature and weak form choices in collocation type discontinuous Galerkin spectral element methods*, J. Sci. Comput., 44 (2010), pp. 136–155.

[27] L. Krivodonova and R. Qin, *An analysis of the spectrum of the discontinuous Galerkin method*, Appl. Numer. Math., 64 (2013), pp. 1–18, https://doi.org/https://doi.org/10.1016/j.apnum.2012.07.008.

[28] D. Kuzmin, *A vertex-based hierarchical slope limiter for p-adaptive discontinuous galerkin methods*, J. Comput. Appl. Math., 233 (2010), pp. 3077–3085, https://doi.org/https://doi.org/10.1016/j.cam.2009.05.028.

[29] R. J. LeVeque, *High-resolution conservative algorithms for advection in incompressible flow*, SIAM J. Numer. Anal., 33 (1996), pp. 627–665, https://doi.org/10.1137/0733033.

[30] D. Ling, J. Cheng, and C.-W. Shu, *Conservative high order positivity-preserving discontinuous galerkin methods for linear hyperbolic and radiative transfer equations.*, J. Sci. Comput., 77 (2018), pp. 1801–1831, https://doi.org/10.1007/s10915-018-0700-3.

[31] H. Liu and J. Qiu, *Finite difference Hermite WENO schemes for hyperbolic conservation laws*, Journal of Scientific Computing, 63 (2015), pp. 548–572.

[32] Y. Liu, C.-W. Shu, and M. Zhang, *Entropy stable high order discontinuous Galerkin methods for ideal compressible MHD on structured meshes*, J. Comput. Phys.s, 354 (2018), pp. 163–178, https://doi.org/https://doi.org/10.1016/j.jcp.2017.10.043.

[33] R. E. Lynch, J. R. Rice, and D. H. Thomas, *Direct solution of partial difference equations by tensor product methods*, Numerische Mathematik, 6 (1964), pp. 185–199.

[34] W. Pazner, *Sparse invariant domain preserving discontinuous Galerkin methods with subcell convex limiting*, Comput. Methods Appl. Mech. Engrg., 382 (2021), p. 113876, https://doi.org/https://doi.org/10.1016/j.cma.2021.113876.

[35] W. Pazner and P.-O. Persson, *Approximate tensor-product preconditioners for very high order discontinuous Galerkin methods*, J. Comput. Phys., 354 (2018), pp. 344–369, https://doi.org/https://doi.org/10.1016/j.jcp.2017.10.030.

[36] P.-O. Persson and J. Peraire, *Newton-gmres preconditioning for discontinuous galerkin discretizations of the navier–stokes equations*, SIAM J. Sci. Comput., 30 (2008), pp. 2709–2733, https://doi.org/https://doi.org/10.1137/070692108.

[37] A. Peyvan, K. Shukla, J. Chan, and G. Karniadakis, *High-order methods for hypersonic flows with strong shocks and real chemistry*, J. Comput. Phys., 490 (2023), p. 112310, https://doi.org/https://doi.org/10.1016/j.jcp.2023.112310.

[38] R. J. Plemmons, *m-matrix characterizations. I–nonsingular m-matrices*, Linear Algebra Appl., 18 (1977), pp. 175–188, https://doi.org/https://doi.org/10.1016/0024-3795(77)90073-8.

[39] T. Qin and C.-W. Shu, *Implicit positivity-preserving high-order discontinuous Galerkin methods for conservation laws*, SIAM J. Sci. Comput., 40 (2018), pp. A81–A107, https://doi.org/10.1137/17M112436X.

[40] F. Renac, *A robust high-order discontinuous Galerkin method with large time steps for the compressible Euler equations*, Commun. Math. Sci., 15 (2017), pp. 813–837, https://doi.org/https://dx.doi.org/10.4310/CMS.2017.v15.n3.a11.

[41] F. Renac, *A robust high-order Lagrange-projection like scheme with large time steps for the isentropic Euler equations*, Numer. Math., 135 (2017), pp. 493–519, https://doi.org/https://doi.org/10.1007/s00211-016-0807-0.

[42] F. Renac, *Entropy stable DGSEM for nonlinear hyperbolic systems in nonconservative form with application to two-phase flows*, J. Comput. Phys., 382 (2019), pp. 1–26, https://doi.org/https://doi.org/10.1016/j.jcp.2018.12.035.

[43] F. Renac, *Entropy stable, robust and high-order DGSEM for the compressible multicomponent Euler equations*, J. Comput. Phys., 445 (2021), p. 110584, https://doi.org/10.1016/j.jcp.2021.110584.

[44] F. Renac, *Maximum principle preserving time implicit DGSEM for nonlinear scalar conservation laws*, arXiv:2406.14317 [math.NA], (2024),

https://arxiv.org/abs/2406.14317.
[45] A. M. Rueda-Ramírez, E. Ferrer, D. A. Kopriva, G. Rubio, and E. Valero, *A statically condensed discontinuous galerkin spectral element method on gauss-lobatto nodes for the compressible navier-stokes equations*, J. Comput. Phys., 426 (2021), p. 109953, https://doi.org/https://doi.org/10.1016/j.jcp.2020.109953.
[46] Y. Saad, *Iterative Methods for Sparse Linear Systems*, Society for Industrial and Applied Mathematics, second ed., 2003, https://doi.org/https://doi.org/10.1137/1.9780898718003.
[47] B. Strand, *Summation by parts for finite difference approximations for d/dx*, J. Comput. Phys., 110 (1994), pp. 47–67, https://doi.org/https://doi.org/10.1006/jcph.1994.1005.
[48] E. Tadmor, *The numerical viscosity of entropy stable schemes for systems of conservation laws. i*, Math. Comp., 49 (1987), pp. 91–103.
[49] J. J. W. van der Vegt, Y. Xia, and Y. Xu, *Positivity preserving limiters for time-implicit higher order accurate discontinuous Galerkin discretizations*, SIAM J. Sci. Comput., 41 (2019), pp. A2037–A2063, https://doi.org/10.1137/18M1227998.
[50] C. F. Van Loan, *The ubiquitous Kronecker product*, J. Comput. Appl. Math., 123 (2000), pp. 85–100, https://doi.org/https://doi.org/10.1016/S0377-0427(00)00393-9.
[51] C. F. Van Loan, *Structured Matrix Problems from Tensors*, Springer International Publishing, Cham, 2016, pp. 1–63, https://doi.org/10.1007/978-3-319-49887-4_1.
[52] M. Waruszewski, J. E. Kozdon, L. C. Wilcox, T. H. Gibson, and F. X. Giraldo, *Entropy stable discontinuous Galerkin methods for balance laws in non-conservative form: Applications to the Euler equations with gravity*, J. Comput. Phys., 468 (2022), p. 111507, https://doi.org/https://doi.org/10.1016/j.jcp.2022.111507.
[53] A. R. Winters, D. Derigs, G. J. Gassner, and S. Walch, *A uniquely defined entropy stable matrix dissipation operator for high Mach number ideal MHD and compressible Euler simulations*, J. Comput. Phys., 332 (2017), pp. 274–289.
[54] A. R. Winters and G. J. Gassner, *Affordable, entropy conserving and entropy stable flux functions for the ideal MHD equations*, J. Comput. Phys., 304 (2016), pp. 72–108.
[55] Z. Xu and C.-W. Shu, *High order conservative positivity-preserving discontinuous Galerkin method for stationary hyperbolic equations*, J. Comput. Phys., 466 (2022), p. 111410, https://doi.org/https://doi.org/10.1016/j.jcp.2022.111410.
[56] Z. Xu and C.-W. Shu, *On the conservation property of positivity-preserving discontinuous Galerkin methods for stationary hyperbolic equations*, J. Comput. Phys., 490 (2023), p. 112304, https://doi.org/https://doi.org/10.1016/j.jcp.2023.112304.
[57] S. T. Zalesak, *Fully multidimensional flux-corrected transport algorithms for fluids*, J. Comput. Phys., 31 (1979), pp. 335–362.
[58] X. Zhang and C.-W. Shu, *On maximum-principle-satisfying high order schemes for scalar conservation laws*, J. Comput. Phys., 229 (2010), pp. 3091–3120.
[59] X. Zhang and C.-W. Shu, *On positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes*, J. Comput. Phys., 229 (2010), pp. 8918–8934.
[60] X. Zhang, Y. Xia, and C.-W. Shu, *Maximum-principle-satisfying and positivity-preserving high order discontinuous Galerkin schemes for conservation laws on triangular meshes*, Journal of Scientific Computing, 50 (2012), pp. 29–62.