On the conservation property of positivity-preserving discontinuous Galerkin methods for stationary hyperbolic equations*

Ziyao Xu† and Chi-Wang Shu‡

Abstract

Recently, there has been a series of works on the positivity-preserving discontinuous Galerkin methods for stationary hyperbolic equations, where the notion of mass conservation follows from a straightforward analogy of that of time-dependent problems, i.e. conserving the mass = preserving cell averages during limiting. Based on such a notion, the implementations and theoretical proofs of positivity-preserving limited methods for stationary equations are unnecessarily complicated and constrained. As will be shown in this paper, in some extreme cases, their convergence could even be problematic. In this work, we clarify a more appropriate definition of mass conservation for limiters applied to stationary hyperbolic equations and establish the genuinely conservative high-order positivity-preserving limited discontinuous Galerkin methods based on this definition. The new methods are able to preserve the positivity of solutions of scalar linear equations and scalar nonlinear equations with invariant wind direction, with much simpler implementations and easier proofs for accuracy and the Lax-Wendroff theorem, compared with the existing methods. Two types of positivity-preserving limiters preserving the local mass of stationary equations are developed to accommodate for the new definition of conservation and their accuracy are investigated. We would like to emphasize that a major advantage of the original DG scheme presented in [24] is a sweeping procedure, which allows for the computation of conservative steady-state solutions explicitly, cell by cell, without iterations, even for nonlinear equations as long as the wind direction is fixed. The main contribution of this paper is to introduce a limiting procedure to enforce positivity without changing the conservative property of this original DG scheme. The good performance of the algorithms for stationary hyperbolic equations and their applications in time-dependent problems are demonstrated by ample numerical tests.

Key Words: stationary hyperbolic equations, mass conservation, Lax-Wendroff theorem, positivity-preserving, high order accuracy, discontinuous Galerkin methods

*Research supported by AFOSR grant FA9550-20-1-0055 and NSF grant DMS-2010107.

†Division of Applied Mathematics, Brown University, Providence, RI 02912, USA. E-mail: ziyao_xu1@brown.edu

‡Division of Applied Mathematics, Brown University, Providence, RI 02912, USA. E-mail: chi-wang_shu@brown.edu

1 Introduction

The hyperbolic balance laws are important tools to investigate the phenomenon of flow and transport. In one space dimension, the scalar hyperbolic balance law is typically written in the form of

$$u_t + f(u)_x = s, \tag{1.1}$$

where u is the balanced quantity, f is the flux function, and s is the source term. In particular, if $s = 0$, the equation is called a hyperbolic conservation law and u is the conserved quantity.

Integrated over the spatial interval $[x_1, x_2]$, the hyperbolic equation (1.1) is transformed to the conservative formulation satisfied by the average of u on $[x_1, x_2]$

$$\frac{d\bar{u}}{dt} + \frac{1}{\Delta x}\left(f(x_2) - f(x_1)\right) = \bar{s}, \tag{1.2}$$

where $\Delta x = x_2 - x_1$, $\bar{u}(t) = \frac{1}{\Delta x}\int_{x_1}^{x_2} u(x,t)dx$, $f(x_i) = f(u(x_i,t)), i = 1,2$, and $\bar{s} = \frac{1}{\Delta x}\int_{x_1}^{x_2} s(x,t)dx$.

Drawn from the formulation (1.2), numerous numerical schemes have been designed for the hyperbolic equation (1.1) in the conservative form

$$\frac{d\bar{u}_j}{dt} + \frac{1}{\Delta x_j}\left(\hat{f}_{j+\frac{1}{2}} - \hat{f}_{j-\frac{1}{2}}\right) = \bar{s}_j, \tag{1.3}$$

under the partition $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}], j = 0, \pm 1, \pm 2, \ldots$, for space, where $\Delta x_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$, $\hat{f}_{j\pm\frac{1}{2}}$ are numerical fluxes at $x_{j\pm\frac{1}{2}}$, $\bar{u}_j$ and $\bar{s}_j$ are cell averages of the numerical solution and the source term on I_j, respectively.

Conservation is of great importance for numerical methods for hyperbolic equations, as it is not only a numerical analogy of the theoretical property of hyperbolic balance laws, but more importantly also the Lax-Wendroff theorem [18], which can be briefly stated as follows,

Theorem 1.1. *Consider a sequence of grids with grid sizes $\Delta x_l, \Delta t_l$ converging to zero as $l \to \infty$, and a sequence of numerical solutions $U_l(x,t), l = 1, 2, \ldots$ computed from a consistent and conservative scheme for a hyperbolic equation on these grids. If U_l converges boundedly a.e. to a function u as $l \to \infty$, then u is a weak solution of the hyperbolic equation.*

Roughly speaking, conservative schemes guarantee correct shock speed determined by the Rankine-Hugoniot jump condition thanks to the mass conservation. To make it clear, we sum the equation (1.3) over the cells $I_j, I_{j+1}, \ldots, I_{j+r}$ to obtain the equation

$$\frac{d}{dt}\int_{x_{j-\frac{1}{2}}}^{x_{j+r+\frac{1}{2}}} u dx + \left(\hat{f}_{j+r+\frac{1}{2}} - \hat{f}_{j-\frac{1}{2}}\right) = \int_{x_{j-\frac{1}{2}}}^{x_{j+r+\frac{1}{2}}} s dx, \tag{1.4}$$

which enforces the correct speed of the shock (if there is a shock in the interval $[x_{j-\frac{1}{2}}, x_{j+r+\frac{1}{2}}]$) since the total mass $\int_{x_{j-\frac{1}{2}}}^{x_{j+r+\frac{1}{2}}} u dx = \int_{x_{j-\frac{1}{2}}}^{\xi} u_l dx + \int_{\xi}^{x_{j+r+\frac{1}{2}}} u_r dx$, where ξ is the location of the shock, and u_l and u_r are the states on the left and right sides of the shock, respectively. On the other hand, non-conservative schemes could produce shocks with totally wrong speed and converge to a spurious solution. A well-known example [20] is the Burgers' equation in the non-conservative form $u_t + uu_x = 0$ discretized by a natural upwinding finite difference scheme $u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x} u_j^n \left(u_j^n - u_{j-1}^n\right)$ with the initial condition $u_j^0 = \mathbf{1}(j < 0)$, where $\mathbf{1}(\cdot)$ is the indicator function. It is easy to check that $u_j^n \equiv u_j^0, \forall n$ for the scheme, which is wrong as the physical solution $u(x,t)$ with the initial condition $u_0(x) = \mathbf{1}(x < 0)$ is $u(x,t) = u_0(x - \frac{1}{2}t)$. For deeper discussions about conservative schemes and their significance for time-dependent hyperbolic equations, one can refer to Chapter 12 in the monograph [20].

The discontinuous Galerkin (DG) method is one of the most popular numerical methods solving hyperbolic equations for its advantages in geometric flexibility, local mass conservation, easiness of parallelization and high order accuracy. The DG method was first proposed in 1973 by Reed et al. [24] to compute the stationary linear transport equation, and first analyzed by Lesaint et al. [19] in 1974. It was later developed into the Runge-Kutta discontinuous Galerkin (RKDG) method in a series of papers by Cockburn et al. [8, 7, 6, 5, 9] for time-dependent nonlinear hyperbolic problems. The classic DG scheme for the hyperbolic equation (1.1) is to find $u \in V$, such that

$$\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u_t v dx - \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} f(u) v_x dx + \hat{f}_{j+\frac{1}{2}} v_{j+\frac{1}{2}}^- - \hat{f}_{j-\frac{1}{2}} v_{j-\frac{1}{2}}^+ = \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} s v dx, \quad \forall v \in V, \tag{1.5}$$

for all j, where V is a piecewise polynomial space and $v_{j+\frac{1}{2}}^{\pm} = \lim_{\epsilon \to 0^+} v(x_{j+\frac{1}{2}} \pm \epsilon)$ denote the right and left limits of v at $x_{j+\frac{1}{2}}$. Taking $v = 1$ on I_j and zero anywhere else in (1.5), we recover the conservative formulation (1.3) satisfied by cell averages. Therefore, the unmodulated DG scheme is conservative for hyperbolic equations.

However, conservation is not the only issue we need to consider for numerical schemes. It is well-known that the scalar hyperbolic conservation laws satisfy the maximum-principle, e.g. its physical solution satisfies $m \le u(x,t) \le M, \forall x \in \mathbb{R}, t > 0$, where $m = \min_{x \in \mathbb{R}} u(x,0)$ and $M = \max_{x \in \mathbb{R}} u(x,0)$. These results hold also for periodic boundary condition and for compactly supported solutions, as well as in higher dimensions. If $m = 0$, the property is also called positivity-preserving. For the hyperbolic balance law (1.1) with $s \ge 0$, the solution is positivity-preserving, provided the initial condition and inflow boundary conditions are nonnegative. It is important to keep the positivity/maximum-principle, besides mass conservation, in

numerical schemes, otherwise the numerical solution is not only physically unacceptable, but also may cause severe robustness issues due to the change of hyperbolicity, or when coupled with other physical systems.

There have been intensive studies on positivity-preserving and maximum-principle-satisfying methods. The genuinely high order maximum-principle-satisfying DG method was proposed in 2010 by Zhang and Shu [35] for scalar hyperbolic equations, and is rapidly developed for different problems ever since, e.g. for the Euler equations [36, 37], Navier-Stokes equations [34], shallow water equations [29], convection-diffusion equations [38, 3], and fluid flow in porous media [13, 4, 31, 12], among others. In addition to the Zhang-Shu framework, there are other techniques for positivity-preserving, such as the flux corrected transport (FCT) technique [1, 14, 15], the invariant domain preserving method [11, 22], and hybrid FV/DGSEM method [27], etc.

The framework of the positivity-preserving DG methods proposed by Zhang and Shu [35, 36] is composed of two parts. The first part is problem-dependent, which is to obtain the solution with provable nonnegative cell averages, probably under certain CFL conditions, from the unmodulated DG scheme. Once the cell averages are guaranteed nonnegative, a scaling limiter, which preserves cell averages and does not destroy the original accuracy of the solution [35], [34], is employed such that the entire solution is modified into nonnegative. It is of great importance for the scaling limiter to preserve cell averages for time-dependent problems. We explain the significance of this principle by an example of the positivity-preserving DG method for (1.1) based on the forward Euler or backward Euler time discretization. The equation satisfied by the cell average of the solution on I_j is given as follows

$$\frac{\bar{u}_j^{n+1}-\bar{u}_j^n}{\Delta t}+\frac{1}{\Delta x_j}\left(\hat{f}^m_{j+\frac{1}{2}}-\hat{f}^m_{j-\frac{1}{2}}\right)=\bar{s}_j^m, \tag{1.6}$$

where n denotes the time level t^n and m is taken as n or $n+1$ in the forward-Euler or backward-Euler time discretization, respectively. We denote the modified solution by $\tilde{u}$ to distinguish it from the unmodulated solution u. Since $\bar{\tilde{u}}_j^n=\bar{u}_j^n$ and $\bar{\tilde{u}}_j^{n+1}=\bar{u}_j^{n+1}$ from the property of the limiter, we have the same equation satisfied by the modified solution:

$$\frac{\bar{\tilde{u}}_j^{n+1}-\bar{\tilde{u}}_j^n}{\Delta t}+\frac{1}{\Delta x_j}\left(\hat{f}^m_{j+\frac{1}{2}}-\hat{f}^m_{j-\frac{1}{2}}\right)=\bar{s}_j^m. \tag{1.7}$$

Thus the Lax-Wendroff theorem and a discrete analogy of (1.4) are satisfied by the modified solution $\tilde{u}$ as well, which guarantees the numerical solution (if it converges) converging to a weak solution with the correct shock speed. This is why preserving cell averages is desired in positivity-preserving/maximum-principle satisfying limiters for time-dependent hyperbolic equations.

Besides time-dependent problems, the stationary hyperbolic equations have also attracted the attention of many researchers. The stationary hyperbolic equations have wide applications in steady-state flow and transport problems. Moreover, they are building blocks of the discrete-ordinate method (DOM) for radiative transfer equations (RTE), see [10, 16]. They are also encountered in implicit time-discretization for time-dependent hyperbolic problems. Similar to the time-dependent problems, the physical solutions of stationary hyperbolic equations are also positivity-preserving, provided the inflow boundary conditions and source terms are nonnegative. There is a series of works on the positivity-preserving DG methods for stationary hyperbolic equations to enhance the stability of numerical algorithms. In 2016, Yuan et al. [32] proposed a rotational limiter based non-conservative positivity-preserving algorithm for constant coefficients stationary hyperbolic equations in one and two space dimensions on structured meshes. Later on, the algorithm is extended to triangular meshes in two space dimensions by Zhang et al. [33] based on a rotational limiter defined on triangles, which is still non-conservative. In 2018, Ling et al. [21] improved the results in [32] in one dimensional space by proving the positivity of cell averages of the unmodulated DG scheme, which results in a high order conservative positivity-preserving DG method by adopting the scaling limiter [35] from time-dependent problems. However, the unmodulated scheme fails to preserve the positivity of cell averages in two space dimensions [21], thus only a second order conservative positivity-preserving scheme was proposed therein by an augmentation of the DG function space. The above works only focus on equations with constant coefficients, and higher than second order conservative methods are unavailable in two and three space dimensions. More recently, we developed high order conservative positivity-preserving algorithms for linear variable coefficient and nonlinear stationary hyperbolic equations in one dimension, and constant coefficients equations in two and three dimensions in [30].

Here, we would like to note that, the notion of conservation in the aforementioned works for stationary hyperbolic equations are different from the notion to be clarified in this paper. The previous notion of conservation in positivity-preserving limiters, coming directly from time-dependent problems to preserve the cell average, is not very suitable for stationary problems.

To show this, we consider the stationary equation

$$f(u)_x + \lambda u = s(x), \tag{1.8}$$

where $f(u)$ is a smooth flux function with unchanged wind direction: $f'(u) > 0, \forall u$, and $\lambda, s(x) \geq 0$ are nonnegative coefficient and source, respectively. The equation (1.8) could come from the backward

Euler discretization of the time-dependent problem (1.1), with the correspondence $u = u^n, \lambda = \frac{1}{\Delta t}$ and $s(x) = \frac{1}{\Delta t}u^{n-1}(x) + s(x, t^n)$. The linear stationary hyperbolic equations, with the main applications in RTE [32, 21, 33], will also be discussed in later sections. Throughout the paper, we always assume the wind direction of the flux in hyperbolic equations does not change, for both nonlinear equations and linear ones, and always use the upwind flux in the DG schemes.

The unmodulated DG scheme with the upwind flux for the equation (1.8) is to find $u \in V$, such that

$$\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \lambda uv dx - \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} f(u)v_x dx + f(u^-_{j+\frac{1}{2}})v^-_{j+\frac{1}{2}} - f(u^-_{j-\frac{1}{2}})v^+_{j-\frac{1}{2}} = \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} svdx, \quad \forall v \in V. \tag{1.9}$$

In the implementation, because of the upwind mechanism of the equation and scheme, we sweep the computation from the left to the right cells, i.e. we obtain the solution u_{j-1} on I_{j-1} before computing u_j on I_j, and then solve u_{j+1} on I_{j+1}, and so forth. Same as the time-dependent cases, by taking $v = 1$ on I_j and zeros on other cells, we obtain the conservation equation satisfied by the cell averages as follows

$$\lambda \Delta x_j \bar{u}_j + f(u^-_{j+\frac{1}{2}}) = f(u^-_{j-\frac{1}{2}}) + \Delta x_j \bar{s}_j,$$

where the right hand side is known when solving u_j. In the positivity-preserving algorithms, the limiter has been used for u_{j-1} on the upstream cell I_{j-1}, to obtain $\tilde{u}_{j-1}(x)$, before the computation of the DG solution u_j, thus the true identity satisfied by the cell average of the DG solution on I_j, before limiting but after the solution on I_{j-1} has been limited, is

$$\lambda \Delta x_j \bar{u}_j + f(u^-_{j+\frac{1}{2}}) = f(\tilde{u}^-_{j-\frac{1}{2}}) + \Delta x_j \bar{s}_j, \tag{1.10}$$

where $\tilde{u}^-_{j-\frac{1}{2}} = \tilde{u}_{j-1}(x_{j-\frac{1}{2}})$ denotes the value of the limiter-modified solution $\tilde{u}_{j-1}(x)$ on I_{j-1} evaluated at $x_{j-\frac{1}{2}}$.

If the limiter is "conservative" in the sense of preserving cell averages, i.e. $\bar{\tilde{u}}_j = \bar{u}_j$, then from (1.10) we have the following equation satisfied by the cell average after the limiter in cell I_j is performed:

$$\lambda \Delta x_j \bar{\tilde{u}}_j + f(\tilde{u}^-_{j+\frac{1}{2}}) = f(\tilde{u}^-_{j-\frac{1}{2}}) + \Delta x_j \bar{s}_j + \left(f(\tilde{u}^-_{j+\frac{1}{2}}) - f(u^-_{j+\frac{1}{2}}) \right), \tag{1.11}$$

Summing the above equations over cells $I_j, I_{j+1}, \ldots, I_{j+r}$ yields

$$\lambda \int_{x_{j-\frac{1}{2}}}^{x_{j+r+\frac{1}{2}}} \tilde{u} dx + f(\tilde{u}^-_{j+r+\frac{1}{2}}) = f(\tilde{u}^-_{j-\frac{1}{2}}) + \int_{x_{j-\frac{1}{2}}}^{x_{j+r+\frac{1}{2}}} s dx + \sum_{i=j}^{j+r} \left(f(\tilde{u}^-_{i+\frac{1}{2}}) - f(u^-_{i+\frac{1}{2}}) \right) \neq f(\tilde{u}^-_{j-\frac{1}{2}}) + \int_{x_{j-\frac{1}{2}}}^{x_{j+r+\frac{1}{2}}} s dx. \tag{1.12}$$

We shall give concrete examples in the numerical section to show that the limiter preserving cell averages for stationary hyperbolic equations could produce solutions with wrong total mass/ shock location, even for the simplest hyperbolic equation $u_t + u_x = 0$ discretized implicitly in time.

On the other hand, if we define the local mass in stationary hyperbolic equations as the sum of the cell average and the outflow flux, and develop limiters such that the modified solution $\tilde{u}$ preserves the local mass on I_j in the sense that

$$\lambda \Delta x_j \bar{\tilde{u}}_j + f(\tilde{u}^-_{j+\frac{1}{2}}) = \lambda \Delta x_j \bar{u}_j + f(u^-_{j+\frac{1}{2}}), \tag{1.13}$$

then we have the local conservation formulation

$$\lambda \Delta x_j \bar{\tilde{u}}_j + f(\tilde{u}^-_{j+\frac{1}{2}}) = f(\tilde{u}^-_{j-\frac{1}{2}}) + \Delta x_j \bar{s}_j, \tag{1.14}$$

and the global conservation formulation

$$\lambda \int_{x_{j-\frac{1}{2}}}^{x_{j+r+\frac{1}{2}}} \tilde{u} dx + f(\tilde{u}^-_{j+r+\frac{1}{2}}) = f(\tilde{u}^-_{j-\frac{1}{2}}) + \int_{x_{j-\frac{1}{2}}}^{x_{j+r+\frac{1}{2}}} s dx, \tag{1.15}$$

satisfied by the modified solution $\tilde{u}$.

In this work, we shall develop two types of slope limiters, named type-1 and type-2, to attain the positivity of the solution, and meanwhile preserve the local mass in the sense of (1.13).

To this end, we would like to give a remark on the definition (1.13) for conservative limiters. Indeed, it is quite reasonable to preserve the sum of the cell average and outflow fluxes in limiters, as any decrease in cell average caused by limiters should be remedied to the mass on the downstream cells via increasing the outflow fluxes in the current cell, and vice versa.

As we will see in later sections, based on this novel definition of conservation, the positivity-preserving DG methods for stationary hyperbolic equations are straightforward and their implementations are simple. We only discuss the linear stationary hyperbolic equations in one and two space dimensions, and nonlinear stationary equations in one dimension to save space, but the method can be directly extended to higher dimensions with various meshes and a class of nonlinear hyperbolic systems with eigenvalues being of the same sign. As important applications, the algorithms developed in this paper can be used in the positivity-preserving algorithm for radiative transfer equations and implicit time discretization for time-dependent hyperbolic problems, see the numerical section and refer to [21] for more details.

We would like to emphasize that a major advantage of the original DG scheme presented in [24] is a sweeping procedure, which allows for the computation of conservative steady-state solutions explicitly, cell

by cell, without iterations, even for nonlinear equations as long as the wind direction is fixed. The main contribution of this paper is to introduce a limiting procedure to enforce positivity without changing the conservative property of this original DG scheme.

The rest of the paper is organized as follows. In Section 2, we establish the positivity-preserving discontinuous Galerkin method for stationary linear hyperbolic equations in one space dimension and construct the conservative limiters with rigorous proofs for the accuracy. We extend the method and limiters to rectangular meshes and triangular meshes in two dimensions in Section 3 and Section 4, respectively. The positivity-preserving technique for stationary nonlinear hyperbolic equations is studied in Section 5, which is focused on one dimension to save space but the method can be extended to higher dimensions directly as in the linear case. In Section 6, we give ample numerical tests to demonstrate the accuracy and effectiveness of our positivity-preserving methods for stationary equations as well as the applications in implicit time discretization for time-dependent problems. Finally, we end up with some concluding remarks in Section 7.

2 Linear stationary hyperbolic equations in one dimension

In this section, we study the high order conservative positivity-preserving discontinuous Galerkin method for the linear stationary hyperbolic equation

$$(a(x)u)_x + \lambda u = s(x), \quad x \in \Omega = (0,1), \tag{2.1}$$

with $0 < a_* \le a(x) \le a^*$ for some positive constants a_*, a^*, and $\lambda, s(x) \ge 0$. We assign the inflow boundary condition $u(0) = u_0 \ge 0$ for the equation. The other case $a(x) < 0$ with boundary condition $u(1) = u_0 \ge 0$ can be transformed to this case by the change of variable $x' = 1 - x$ in (2.1), thus we omit the discussion. We assume λ is constant for simplicity, as we are mainly concerned with the applications of the model in the discrete-ordinate method (DOM) for radiative transfer equations (RTE) and implicit time-discretization for time-dependent hyperbolic problems, where λ is constant for both cases. However, there is not essential difficulty to extend the positivity-preserving technique to the variable case $\lambda(x) \ge 0$.

We adopt the partition $0 = x_{\frac{1}{2}} < x_{\frac{3}{2}} < \cdots < x_{N+\frac{1}{2}} = 1$ for Ω and denote the j-th cell by $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ with the length $\Delta x_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$, for $j = 1, 2, \ldots, N$. The function space V of the P^k-DG scheme is defined as

$$V = \left\{ v \in L^2(\Omega) : v|_{I_j} \in P^k(I_j), j = 1, 2, \ldots, N \right\},$$

where $P^k(I_j)$ denotes the space of polynomials of order no greater than k on the cell I_j. We define the cell average of $v \in V$ on I_j as $\bar{v}_j = \frac{1}{\Delta x_j}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} v(x)dx$, and its left and right limits at the interface $x_{j+\frac{1}{2}}$ as $v^{\pm}_{j+\frac{1}{2}} = v(x_{j+\frac{1}{2}} \pm 0)$. Moreover, we denote by $v_j = v|_{I_j}$ for $v \in V$, $j = 1, 2, \ldots, N$, for convenience.

The positivity-preserving P^k-DG scheme of the equation (2.1) is to find $u \in V$, such that

$$-\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} a(x)uv_x dx + a(x_{j+\frac{1}{2}})u^-_{j+\frac{1}{2}}v^-_{j+\frac{1}{2}} + \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \lambda uv dx = a(x_{j-\frac{1}{2}})\tilde{u}^-_{j-\frac{1}{2}}v^+_{j-\frac{1}{2}} + \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} svdx, \quad \forall v \in P^k(I_j) \tag{2.2}$$

for $j = 1, 2, \ldots, N$, where we define $u^-_{\frac{1}{2}} = u_0$. We would like to emphasize that, the calculation of u_j is based on the modified solution on the upstream cells, thus we use $\tilde{u}_{j-1}$ on the right hand side of the scheme (2.2). Once u_j is solved from the scheme, we employ the positivity-preserving limiter to be introduced later to obtain the modified solution $\tilde{u}_j$, and use it in the calculation of u_{j+1}, and so forth.

Assume the quadrature rules adopted in the scheme (2.2) is accurate for integrals of k-th order polynomials. Taking the test function $v = 1$ on I_j in the scheme (2.2), we obtain the following equation satisfied by the local mass

$$\lambda \Delta x_j \bar{u}_j + a(x_{j+\frac{1}{2}})u^-_{j+\frac{1}{2}} = a(x_{j-\frac{1}{2}})\tilde{u}^-_{j-\frac{1}{2}} + \Delta x_j \bar{s}_j, \tag{2.3}$$

For convenience, we define $LHS(w_j) = \lambda \Delta x_j \bar{w}_j + a(x_{j+\frac{1}{2}})w^-_{j+\frac{1}{2}}$, for $w_j \in P^k(I_j)$, to be the amount of local mass of w_j on I_j. Since $\tilde{u}^-_{j-\frac{1}{2}} \geq 0$ on the right hand side of (2.3), we have $LHS(u_j) \geq 0$. The conservative limiter should satisfy $LHS(\tilde{u}_j) = LHS(u_j)$, where u_j and $\tilde{u}_j$ are the unmodulated and modified solutions on I_j, respectively.

There are two types of limiters to be developed throughout the paper, where the type-1 limiter requires the DG scheme to use the Gauss-Radau quadrature rule of $k+1$ points for numerical integration and only guarantees the positivity of modified solution at the Gauss-Radau points (it may be negative at some other points, while the type-2 limiter does not have such requirement. We denote the Gauss-Radau points on I_j by $\hat{x}_\alpha, \alpha = 1, 2, \ldots, k+1$ with $\hat{x}_{k+1} = x_{j+\frac{1}{2}}$, and the corresponding weights by $\hat{\omega}_\alpha, \alpha = 1, 2, \ldots, k+1$ with $\sum_{\alpha=1}^{k+1} \hat{\omega}_\alpha = 1$.

The type-1 limiter for u_j is defined as follows:

$$\tilde{u}_j(x) = \theta_j \hat{u}_j(x), \quad \hat{u}_j(x) = \sum_{\alpha=1}^{k+1} u_j^+(\hat{x}_\alpha)\ell_\alpha(x), \tag{2.4}$$

where $z^+ = \max\{z, 0\}$ is the positive part of a real number z, $\ell_\alpha(x)$ is the Lagrange basis at the Gauss-Radau points $\{\hat{x}_\beta\}_{\beta=1}^{k+1}$ with $\ell_\alpha(\hat{x}_\beta) = \delta_{\alpha,\beta}$, and $\theta_j = \frac{LHS(u_j)}{LHS(\hat{u}_j)} \in [0, 1]$. Note that the integral in $LHS(\cdot)$ is evaluated

by the Gauss-Radau quadrature, thus $0 \le LHS(u_j) \le LHS(\hat{u}_j)$. In the case $LHS(u_j) = LHS(\hat{u}_j) = 0$, we take $\theta_j = 1$. In practice, this case can be avoided by taking $\theta_j = \frac{LHS(u_j)+\varepsilon}{LHS(\hat{u}_j)+\varepsilon}$, where ε is a very small positive number, e.g. $\varepsilon = 10^{-16}$.

It is clear that the limiter (2.4) is conservative in the sense that $LHS(\tilde{u}_j) = LHS(u_j)$, and $\tilde{u}_j \ge 0$ at the Gauss-Radau points $\{\hat{x}_\alpha\}_{\alpha=1}^{k+1}$. More importantly, we have the result of accuracy for the limiter as follows:

Lemma 2.1. *Consider the solution u_j of the scheme (2.2) with an L^∞-error of $O(\Delta x_j^{k+1})$. If $\lambda = 0$, the error introduced by the limiter (2.4) is $||\tilde{u}_j - u_j||_{L^\infty(I_j)} = O(\Delta x_j^{k+1})$. If $\lambda > 0$, the error introduced by the limiter (2.4) is $||\tilde{u}_j - u_j||_{L^\infty(I_j)} = O(\Delta x_j^{k})$, but the error is optimal at the downstream point, i.e. $|\tilde{u}^-_{j+\frac{1}{2}} - u^-_{j+\frac{1}{2}}| = O(\Delta x_j^{k+1})$.*

Proof. We can decompose the error as

$$e = u_j - \tilde{u}_j = (u_j - \hat{u}_j) + (\hat{u}_j - \tilde{u}_j) = e_1 + e_2. \tag{2.5}$$

For $e_1 = u_j - \hat{u}_j$, we have the estimate

$$\begin{aligned} |e_1(x)| &= |\hat{u}_j(x) - u_j(x)| \\ &= \left| \sum_{\alpha=1}^{k+1} u_j^+(\hat{x}_\alpha)\ell_\alpha(x) - \sum_{\alpha=1}^{k+1} u_j(\hat{x}_\alpha)\ell_\alpha(x) \right| \\ &= \left| \sum_{\alpha=1}^{k+1} u_j^-(\hat{x}_\alpha)\ell_\alpha(x) \right| \\ &\le \sum_{\alpha=1}^{k+1} |\ell_\alpha(x)| \cdot \max_{1\le\alpha\le k+1} u_j^-(\hat{x}_\alpha) \\ &\le \Lambda_k \cdot O(\Delta x_j^{k+1}) = O(\Delta x_j^{k+1}), \quad \forall x \in I_j, \end{aligned} \tag{2.6}$$

where $z^- = -\min\{z, 0\}$ denotes the negative part of a real number z and $\Lambda_k = \max_{x\in I_j} \sum_{\alpha=1}^{k+1} |\ell_\alpha(x)|$ is the Lebesgue constant. Note that $u_j^-(\hat{x}_\alpha) = O(\Delta x_j^{k+1}), \alpha = 1, \ldots, k+1$, since the exact solution is nonnegative. Therefore $||e_1||_{L^\infty(I_j)} = O(\Delta x_j^{k+1})$.

For e_2, we have $e_2 = \hat{u}_j - \tilde{u}_j = (1-\theta_j)\hat{u}_j$. If $\lambda = 0$, we have $e_2 \equiv 0$ since $\theta_j = 1$, which follows from the

observation that $u^-_{j+\frac{1}{2}} = \hat{u}^-_{j+\frac{1}{2}} \geq 0$ due to (2.2), (2.3). If $\lambda > 0$, we have the estimate for $e_2(x)$ as follows,

$$\begin{aligned}
|e_2(x)| &= (1-\theta_j)\,|\hat{u}_j(x)| = \left(1 - \frac{LHS(u_j)}{LHS(\hat{u}_j)}\right)|\hat{u}_j(x)| = \frac{LHS(\hat{u}_j - u_j)}{LHS(\hat{u}_j)}\,|\hat{u}_j(x)| \\
&= \frac{\lambda\Delta x_j\left(\bar{\hat{u}}_j - \bar{u}_j\right) + a(x_{j+\frac{1}{2}})\left(\hat{u}^-_{j+\frac{1}{2}} - u^-_{j+\frac{1}{2}}\right)}{\lambda\Delta x_j\bar{\hat{u}}_j + a(x_{j+\frac{1}{2}})\hat{u}^-_{j+\frac{1}{2}}}\,|\hat{u}_j(x)| \qquad . \qquad (2.7)\\
&\leq \frac{\lambda\Delta x_j||e_1||_{L^\infty(I_j)} + a(x_{j+\frac{1}{2}})||e_1||_{L^\infty(I_j)}}{\lambda\Delta x_j\bar{\hat{u}}_j + a(x_{j+\frac{1}{2}})\hat{u}^-_{j+\frac{1}{2}}}\,|\hat{u}_j(x)|
\end{aligned}$$

In particular, at the Gauss-Radau points, we have the following estimates from (2.7),

$$\begin{aligned}
|e_2(\hat{x}_\beta)| &\leq \frac{\lambda\Delta x_j||e_1||_{L^\infty(I_j)} + a(x_{j+\frac{1}{2}})||e_1||_{L^\infty(I_j)}}{\lambda\Delta x_j\bar{\hat{u}}_j}\hat{u}_j(\hat{x}_\beta) \\
&= \frac{\lambda\Delta x_j||e_1||_{L^\infty(I_j)} + a(x_{j+\frac{1}{2}})||e_1||_{L^\infty(I_j)}}{\lambda\Delta x_j\sum_{\alpha=1}^{k+1}\hat{\omega}_\alpha\hat{u}_j(\hat{x}_\alpha)}\hat{u}_j(\hat{x}_\beta) \\
&\leq \frac{\lambda\Delta x_j||e_1||_{L^\infty(I_j)} + a(x_{j+\frac{1}{2}})||e_1||_{L^\infty(I_j)}}{\lambda\Delta x_j\hat{\omega}_\beta\hat{u}_j(\hat{x}_\beta)}\hat{u}_j(\hat{x}_\beta) \qquad (2.8)\\
&\leq \left(\hat{\omega}_\beta^{-1} + \frac{a^*}{\lambda\hat{\omega}_\beta\Delta x_j}\right)||e_1||_{L^\infty(I_j)} \\
&= O(\Delta x_j^k), \quad \text{for } \beta = 1, 2, \ldots, k+1,
\end{aligned}$$

thus,

$$|e_2(x)| = |\sum_{\alpha=1}^{k+1} e_2(\hat{x}_\alpha)\ell_\alpha(x)| \leq \sum_{\alpha=1}^{k+1}|\ell_\alpha(x)| \cdot \max_{1\leq\alpha\leq k+1}|e_2(\hat{x}_\alpha)| \leq \Lambda_k \cdot O(\Delta x_j^k) = O(\Delta x_j^k), \quad \forall x \in I_j, \qquad (2.9)$$

i.e. $||e_2||_{L^\infty(I_j)} = O(\Delta x_j^k)$.

In particular, at the downstream point $x_{j+\frac{1}{2}}$, it follows from (2.7) that

$$\begin{aligned}
|e_2(x_{j+\frac{1}{2}})| &\leq \frac{\lambda\Delta x_j||e_1||_{L^\infty(I_j)} + a(x_{j+\frac{1}{2}})||e_1||_{L^\infty(I_j)}}{\lambda\Delta x_j\sum_{\alpha=1}^{k+1}\hat{\omega}_\alpha\hat{u}_j(\hat{x}_\alpha) + a(x_{j+\frac{1}{2}})\hat{u}^-_{j+\frac{1}{2}}}\hat{u}^-_{j+\frac{1}{2}} \\
&= \frac{\lambda\Delta x_j||e_1||_{L^\infty(I_j)}}{\lambda\Delta x_j\sum_{\alpha=1}^{k+1}\hat{\omega}_\alpha\hat{u}_j(\hat{x}_\alpha) + a(x_{j+\frac{1}{2}})\hat{u}^-_{j+\frac{1}{2}}}\hat{u}^-_{j+\frac{1}{2}} + \frac{a(x_{j+\frac{1}{2}})||e_1||_{L^\infty(I_j)}}{\lambda\Delta x_j\sum_{\alpha=1}^{k+1}\hat{\omega}_\alpha\hat{u}_j(\hat{x}_\alpha) + a(x_{j+\frac{1}{2}})\hat{u}^-_{j+\frac{1}{2}}}\hat{u}^-_{j+\frac{1}{2}} \\
&\leq \frac{\Delta x_j||e_1||_{L^\infty(I_j)}}{\Delta x_j\hat{\omega}_{k+1}\hat{u}^-_{j+\frac{1}{2}}}\hat{u}^-_{j+\frac{1}{2}} + \frac{a(x_{j+\frac{1}{2}})||e_1||_{L^\infty(I_j)}}{a(x_{j+\frac{1}{2}})\hat{u}^-_{j+\frac{1}{2}}}\hat{u}^-_{j+\frac{1}{2}} \\
&\leq \left(1 + \hat{\omega}_{k+1}^{-1}\right)||e_1||_{L^\infty(I_j)} = O(\Delta x_j^{k+1}).
\end{aligned} \qquad (2.10)$$

Gathering all results above and using the triangle inequalities, we finish the proof of Lemma 2.1. □

We would like to note that, the error estimates in Lemma 2.1 is sharp and the result cannot be improved by any conservative limiters, which can be illustrated by a concrete example given as follows.

Example 2.1. *We assume $\lambda > 0$. Consider the numerical approximation $u_j(x) = x_{j+\frac{1}{2}} - x - \Delta x_j^{k+1}$ of the exact solution $u_j^{exact}(x) = x_{j+\frac{1}{2}} - x$ on the cell $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$. The modified solution $\tilde{u}_j$ of any conservative limiters should satisfy $\lambda\Delta x_j\bar{\tilde{u}}_j + a(x_{j+\frac{1}{2}})\tilde{u}^-_{j+\frac{1}{2}} = \lambda\Delta x_j\bar{u}_j + a(x_{j+\frac{1}{2}})u^-_{j+\frac{1}{2}}$. Since $u^-_{j+\frac{1}{2}} = -\Delta x_j^{k+1}$ and $\tilde{u}^-_{j+\frac{1}{2}} \geq 0$, we have $\bar{u}_j - \bar{\tilde{u}}_j = \frac{a(x_{j+\frac{1}{2}})}{\lambda\Delta x_j}\left(\tilde{u}^-_{j+\frac{1}{2}} - u^-_{j+\frac{1}{2}}\right) \geq \frac{a(x_{j+\frac{1}{2}})}{\lambda\Delta x_j}\left(0 - (-\Delta x_j^{k+1})\right) = a(x_{j+\frac{1}{2}})\lambda^{-1}\Delta x_j^k$, which implies that $\tilde{u}_j$ is at most k-th order accurate.*

The type-1 limiter only preserves the positivity of modified solutions at the Gauss-Radau points and we must use the Gauss-Radau quadrature to evaluate integrals in the scheme (2.2), which may not be satisfactory in some applications. We can define a type-1 limiter that preserves positivity at the $k+1$ Gauss-Lobatto points by (2.4) as well. However, the Gauss-Lobatto quadrature has lower algebraic accuracy than the Gauss-Radau quadrature with the same number of points and may lead to possible loss of accuracy up to one order [2], hence we do not consider it in this paper. We now introduce the type-2 limiter, which is positivity-preserving on the whole cell or at any desired points, and exempts the requirement on quadrature rules.

The type-2 limiter is defined as follows,

$$\tilde{u}_j(x) = \theta_j\hat{u}_j(x), \quad \hat{u}_j(x) = u_j(x) + \epsilon_j, \tag{2.11}$$

where $\epsilon_j = -\min\{\min_{x\in S} u_j(x), 0\}$, $S \subset I_j$ is the set of points where we want to preserve the positivity of the solution, and $\theta_j = \frac{LHS(u_j)}{LHS(\hat{u}_j)} \in [0, 1]$.

It is clear that the limiter (2.11) is conservative in the sense that $LHS(\tilde{u}_j) = LHS(u_j)$ and $\tilde{u}_j \geq 0$ on S. More importantly, we have the accuracy result for the limiter as follows:

Lemma 2.2. *Consider the solution u_j of the scheme (2.2) with an L^∞-error of $O(\Delta x_j^{k+1})$. If $\lambda = 0$, the error introduced by the limiter (2.11) is $||\tilde{u}_j - u_j||_{L^\infty(I_j)} = O(\Delta x_j)$. If $\lambda > 0$, the error introduced by the limiter (2.11) is $||\tilde{u}_j - u_j||_{L^\infty(I_j)} = O(\Delta x_j^k)$. Nevertheless, at the downstream point, the errors in both cases are optimal, i.e. $|\tilde{u}^-_{j+\frac{1}{2}} - u^-_{j+\frac{1}{2}}| = O(\Delta x_j^{k+1})$.*

Proof. For simplicity, we assume $S = I_j$. We have the same decomposition $e = u_j - \tilde{u}_j = (u_j - \hat{u}_j) + (\hat{u}_j - \tilde{u}_j) = e_1 + e_2$ for the error as (2.5).

It is clear that $||e_1||_{L^\infty(I_j)} = ||u_j - \hat{u}_j||_{L^\infty(I_j)} = \epsilon_j = O(\Delta x_j^{k+1})$ by the definitions. For e_2, we have $e_2 = \hat{u}_j - \tilde{u}_j = (1-\theta_j)\hat{u}_j$. If $\lambda = 0$, we have $||e_2||_{L^\infty(I_j)} = (1-\theta_j)||\hat{u}_j||_{L^\infty(I_j)} = O(\Delta x_j)$, since $||\hat{u}_j||_{L^\infty(I_j)} = O(\Delta x_j)$ due to $\hat{u}_j(\xi_j) = 0$ at its minimum point $\xi_j \in I_j$ when $\theta_j < 1$. If $\lambda > 0$, we have the estimates for e_2 exactly the same as (2.7), (2.8), (2.9), and end up with the result $||e_2||_{L^\infty(I_j)} = O(\Delta x_j^k)$.

At the downstream point $x_{j+\frac{1}{2}}$, the estimate for e_2 is exactly the same as (2.10), thereby $|e_2(x_{j+\frac{1}{2}})| = O(\Delta x_j^{k+1})$.

Gathering all results above and using the triangle inequalities, we finish the proof of Lemma 2.2. □

The estimates in Lemma (2.2) is sharp, i.e. it could happen that $||\tilde{u}_j - u_j||_{L^\infty(I_j)} = O(\Delta x_j)$ if $\lambda = 0$, which can be illustrated by the following example.

Example 2.2. *We assume $\lambda = 0$. Consider the exact solution u_j^{exact} on I_j with $u_j^{exact}(\hat{x}_\alpha) = \Delta x_j$ for $1 \le \alpha \le k-1$ and $u_j^{exact}(\hat{x}_\alpha) = 0$ for $\alpha = k, k+1$, and its numerical approximation $u_j = \sum_{\alpha=1}^{k+1} u_j^{exact}(\hat{x}_\alpha)\ell_\alpha(x) - \Delta x_j^{k+1}\ell_k(x)$. A graph of u_j in the case $k = 2$ is given in Figure 1.*

It is clear that $LHS(u_j) = 0$, since $\lambda = 0$ and $u_j(x_{j+\frac{1}{2}}) = 0$. On the other hand, we have $\epsilon_j > 0$ in the limiter (2.11), since $u_j(\hat{x}_k) = -\Delta x_j^{k+1} < 0$. Thus we can compute θ_j by definition: $\theta_j = \frac{LHS(u_j)}{LHS(\hat{u}_j)} = \frac{LHS(u_j)}{LHS(u_j+\epsilon_j)} = \frac{LHS(u_j)}{LHS(u_j)+LHS(\epsilon_j)} = \frac{0}{0+a(x_{j+\frac{1}{2}})\epsilon} = 0$, which implies $\tilde{u}_j = \theta_j\hat{u}_j = 0$. Since u_j is flattened to $\tilde{u}_j \equiv 0$, the limiter (2.11) is only of the accuracy $O(\Delta x)$ in this case.

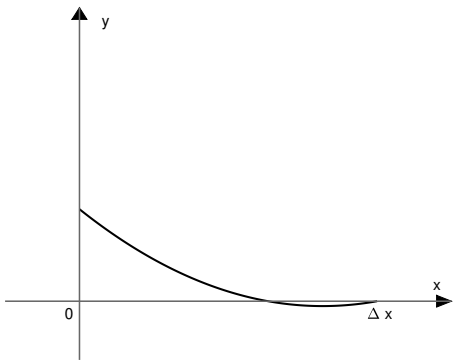


Figure 1: A graph of quadratic polynomial u_j in Example 2.2

Remark 2.1. *The above discussions are based on the assumption that λ is a constant. However, in the backward Euler discretization for time-dependent problems, λ is of the order $\frac{1}{\Delta t}$, as demonstrated in the introduction. If we take the common CFL condition $\Delta t \propto \Delta x$ in this case, the accuracy of both the type-1 and type-2 limiters is optimal, which is clear from the estimates in the proofs. The same conclusion applies to later sections.*

Since the accuracy of both type-1 and type-2 limiters is optimal at the downstream points of cells, the

possible non-optimal errors introduced by the limiters do not propagate to downstream cells, which makes the limited positivity-preserving DG solution having the optimal order of accuracy in the sense of downstream points of cells.

Collecting the Lemmas 2.1 and 2.2, we attain the following theorem for the positivity-preserving DG method of the equation (2.1).

Theorem 2.3. *For the linear stationary hyperbolic equation* (2.1)*, if the source term and inflow boundary condition are nonnegative, then the solution of the scheme* (2.2) *modified by the limiter* (2.4) *or* (2.11) *is nonnegative, with the local accuracy established in Lemmas 2.1 and 2.2, respectively.*

3 Linear stationary hyperbolic equations in two dimensions on rectangular meshes

In this section, we study the high order conservative positivity-preserving discontinuous Galerkin method in two space dimensions on rectangular meshes for the linear stationary hyperbolic equation

$$(a(x,y)u)_x + (b(x,y)u)_y + \lambda u = s(x,y), \quad (x,y) \in \Omega = (0,1)^2, \tag{3.1}$$

with $0 < a_* \le a(x,y) \le a^*$ and $0 < b_* \le b(x,y) \le b^*$ for some positive constants a_*, a^*, b_*, b^*, and $\lambda, s(x,y) \ge 0$. We assign the inflow boundary conditions $u(x,0) = g_1(x) \ge 0$ and $u(0,y) = g_2(y) \ge 0$ for the equation. The cases $a(x,y) < 0$ and/or $b(x,y) < 0$ can be transformed to this case by the change of variables $x' = 1 - x$ and/or $y' = 1 - y$, thus we omit the discussion.

We partition the domain Ω by $0 \le x_{\frac{1}{2}} < x_{\frac{3}{2}} < \ldots < x_{N_x+\frac{1}{2}} = 1$ and $0 \le y_{\frac{1}{2}} < y_{\frac{3}{2}} < \ldots < y_{N_y+\frac{1}{2}} = 1$ in the x and y directions, respectively, and denote by $K_{i,j} = I_i \times J_j = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$ the cells in Ω with the area $|K_{i,j}| = \Delta x_i \Delta y_j$, where $\Delta x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}, \Delta y_j = y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}}$, $i = 1, 2, \ldots N_x, j = 1, 2, \ldots, N_y$. Moreover, we assume the meshes are regular in the refinement, i.e. $\max_{i,j}\{\Delta x_i, \Delta y_j\} \le \rho \min_{i,j}\{\Delta x_i, \Delta y_j\}$ for some constant ρ that is independent of mesh sizes, and denote by $h = \min_{i,j}\{\Delta x_i, \Delta y_j\}$. The function space V of the Q^k-DG scheme is defined as

$$V = \left\{ v \in L^2(\Omega) : v|_{K_{i,j}} \in Q^k(K_{i,j}), i = 1, 2, \ldots, N_x, j = 1, 2, \ldots, N_y \right\},$$

where $Q^k(K)$ denotes the space of tensor products of polynomials of order no greater than k on the cell K.

Similar to the one space dimension, we define the cell average of $v \in V$ on $K_{i,j}$ as $\bar{v}_{i,j} = \frac{1}{\Delta x_i \Delta y_j} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} v(x,y)dxdy$, and its left/right and lower/upper limits on the vertical and horizontal cell interfaces by $v(x^{\pm}_{i+\frac{1}{2}}, y) = \lim_{\epsilon \to 0^+} v(x_{i+\frac{1}{2}} \pm \epsilon, y)$ and $v(x, y^{\pm}_{j+\frac{1}{2}}) = \lim_{\epsilon \to 0^+} v(x, y_{j+\frac{1}{2}} \pm \epsilon)$, respectively. Moreover, we denote by $v_{i,j} = v|_{K_{i,j}}$ for $v \in V, i = 1, 2, \ldots, N_x, j = 1, 2, \ldots, N_y$.

The positivity-preserving Q^k-DG scheme of the equation (3.1) on rectangular meshes is to find $u \in V$, such that

$$
\begin{aligned}
&-\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} (auv_x + buv_y - \lambda uv)\, dxdy + \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} au(x^-_{i+\frac{1}{2}}, y)v(x^-_{i+\frac{1}{2}}, y)dy + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} bu(x, y^-_{j+\frac{1}{2}})v(x, y^-_{j+\frac{1}{2}})dx \\
&= \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} a\tilde{u}(x^-_{i-\frac{1}{2}}, y)v(x^+_{i-\frac{1}{2}}, y)dy + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} b\tilde{u}(x, y^-_{j-\frac{1}{2}})v(x, y^+_{j-\frac{1}{2}})dx + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} svdxdy, \quad \forall v \in Q^k(K_{i,j}),
\end{aligned}
\tag{3.2}
$$

for $i = 1, 2, \ldots, N_x, j = 1, 2, \ldots, N_y$, where we define $u(x, y^-_{\frac{1}{2}}) = \mathcal{I}(g_1)(x)$ and $u(x^-_{\frac{1}{2}}, y) = \mathcal{I}(g_2)(y)$ on the inflow boundaries, with $\mathcal{I}$ denoting the polynomial interpolation at the quadrature points on cell interfaces. In the computation, we solve $u_{i,j}$ on cell $K_{i,j}$ based on the modified solutions $\tilde{u}_{i-1,j}$ and $\tilde{u}_{i,j-1}$ on upstream cells. Once $u_{i,j}$ is obtained, we employ the positivity-preserving limiters to get the modified solution $\tilde{u}_{i,j}$ and use it in the computations on the downstream cells.

Taking the test function $v = 1$ on $K_{i,j}$ in the scheme (3.2), we obtain the following equation satisfied by the local mass

$$
\lambda \Delta x_i \Delta y_j \bar{u}_{i,j} + \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} au(x^-_{i+\frac{1}{2}}, y)dy + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} bu(x, y^-_{j+\frac{1}{2}})dx = \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} a\tilde{u}(x^-_{i-\frac{1}{2}}, y)dy + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} b\tilde{u}(x, y^-_{j-\frac{1}{2}})dx + \Delta x_i \Delta y_j \bar{s}_{i,j}.
\tag{3.3}
$$

We define $LHS(w_{i,j}) = \lambda \Delta x_i \Delta y_j \bar{w}_{i,j} + \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} aw(x^-_{i+\frac{1}{2}}, y)dy + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} bw(x, y^-_{j+\frac{1}{2}})dx$, for $w_{i,j} \in Q^k(K_{i,j})$, to be the amount of local mass of $w_{i,j}$ on $K_{i,j}$. Moreover, we define $LHS^b(w_{i,j}) = \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} aw(x^-_{i+\frac{1}{2}}, y)dy + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} bw(x, y^-_{j+\frac{1}{2}})dx$ for the total outflow flux. Since $\tilde{u}_{i-1,j}, \tilde{u}_{i,j-1} \geq 0$ on the right hand side of (3.3), we have $LHS(u_{i,j}) \geq 0$. In particular, if $\lambda = 0$, then $LHS^b(u_{i,j}) = LHS(u_{i,j}) \geq 0$.

Similar to the one dimensional case, there are two types of limiters, in which the type-1 limiter depends on the Gauss-Radau quadrature while the type-2 limiter does not.

The type-1 limiter for $u_{i,j}$ is defined as follows:

$$\begin{aligned}
&\tilde{u}_{i,j}(x,y) = \theta^2_{i,j}\hat{u}_{i,j}(x,y), \quad \hat{u}_{i,j}(x,y) = \hat{u}^o_{i,j}(x,y) + \theta^1_{i,j}\hat{u}^b_{i,j}(x,y),\\
&\hat{u}^o_{i,j}(x,y) = \sum_{\alpha=1}^{k}\sum_{\beta=1}^{k} u^+_{i,j}(\hat{x}_\alpha,\hat{y}_\beta)\ell_\alpha(x)\ell_\beta(y),\\
&\hat{u}^b_{i,j}(x,y) = \sum_{\alpha=1}^{k} u^+_{i,j}(\hat{x}_\alpha,y^-_{j+\frac{1}{2}})\ell_\alpha(x)\ell_{k+1}(y) + \sum_{\beta=1}^{k+1} u^+_{i,j}(x^-_{i+\frac{1}{2}},\hat{y}_\beta)\ell_{k+1}(x)\ell_\beta(y),
\end{aligned} \tag{3.4}$$

where $\{\hat{x}_\alpha\}_{\alpha=1}^{k+1}$ and $\{\hat{y}_\beta\}_{\beta=1}^{k+1}$ are the Gauss-Radau points on the intervals $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ and $[y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$, respectively, with $\hat{x}_{k+1} = x_{i+\frac{1}{2}}$ and $\hat{y}_{k+1} = y_{j+\frac{1}{2}}$, we abuse notations to denote by $\ell_\alpha(x)$ and $\ell_\beta(y)$ the Lagrange basis at $\{\hat{x}_\alpha\}_{\alpha=1}^{k+1}$ and $\{\hat{y}_\beta\}_{\beta=1}^{k+1}$, respectively, $\theta^1_{i,j} = \max\{\frac{LHS^b(u^b_{i,j})}{LHS^b(\hat{u}^b_{i,j})}, 0\} \in [0,1]$, $\theta^2_{i,j} = \frac{LHS(u_{i,j})}{LHS(\hat{u}_{i,j})} \in [0,1]$, $u^b_{i,j}(x,y) = \sum_{\alpha=1}^{k} u_{i,j}(\hat{x}_\alpha, y^-_{j+\frac{1}{2}})\ell_\alpha(x)\ell_{k+1}(y) + \sum_{\beta=1}^{k+1} u_{i,j}(x^-_{i+\frac{1}{2}},\hat{y}_\beta)\ell_{k+1}(x)\ell_\beta(y)$. In particular, if $\lambda = 0$, we have $\theta^1_{i,j} = \frac{LHS^b(u^b_{i,j})}{LHS^b(\hat{u}^b_{i,j})}$ and $\theta^2_{i,j} = 1$. We denote $u^o_{i,j}(x,y) = \sum_{\alpha=1}^{k}\sum_{\beta=1}^{k} u_{i,j}(\hat{x}_\alpha,\hat{y}_\beta)\ell_\alpha(x)\ell_\beta(y)$ for the convenience of later discussion.

We have the accuracy results for the conservative positivity-preserving limiter (3.4) as follows:

Lemma 3.1. *Consider the solution $u_{i,j}$ of the scheme (3.2) with an L^∞-error of $O(h^{k+1})$. If $\lambda = 0$, the error introduced by the limiter (3.4) is $||\tilde{u}_{i,j} - u_{i,j}||_{L^\infty(K_{i,j})} = O(h^{k+1})$. If $\lambda > 0$, the error introduced by the limiter (3.4) is $||\tilde{u}_{i,j} - u_{i,j}||_{L^\infty(K_{i,j})} = O(h^k)$, but the error is optimal on the downstream edges, i.e. $||\tilde{u}_{i,j} - u_{i,j}||_{L^\infty(I_{i+\frac{1}{2}}\cup J_{j+\frac{1}{2}})} = O(h^{k+1})$, where $I_{i+\frac{1}{2}}$ and $J_{j+\frac{1}{2}}$ denote the right and upper edges of $K_{i,j}$, respectively.*

Proof. We decompose the error as

$$e = u_{i,j} - \tilde{u}_{i,j} = (u_{i,j} - \hat{u}_{i,j}) + (\hat{u}_{i,j} - \tilde{u}_{i,j}) = e_1 + e_2, \tag{3.5}$$

and

$$e_1 = u_{i,j} - \hat{u}_{i,j} = \left(u^o_{i,j} - \hat{u}^o_{i,j}\right) + \left(u^b_{i,j} - \hat{u}^b_{i,j}\right) + \left(\hat{u}^b_{i,j} - \theta^1_{i,j}\hat{u}^b_{i,j}\right) = e_{1,1} + e_{1,2} + e_{1,3}. \tag{3.6}$$

Using similar arguments as in (2.6), it is easy to prove that $||e_{1,1}||_{L^\infty(K_{i,j})} = O(h^{k+1})$ and $||e_{1,2}||_{L^\infty(K_{i,j})} = O(h^{k+1})$. As for $e_{1,3} = \left(1 - \theta^1_{i,j}\right)\hat{u}^b_{i,j}$, we consider two cases.

Case I: $\theta^1_{i,j} = 0$. We have $LHS^b(u^b_{i,j}) \le 0$, i.e.

$$\Delta y_j \sum_{\beta=1}^{k+1} \hat{\omega}_\beta a(x_{i+\frac{1}{2}},\hat{y}_\beta)u_{i,j}(x^-_{i+\frac{1}{2}},\hat{y}_\beta) + \Delta x_i \sum_{\alpha=1}^{k+1} \hat{\omega}_\alpha b(\hat{x}_\alpha, y_{j+\frac{1}{2}})u_{i,j}(\hat{x}_\alpha, y^-_{j+\frac{1}{2}}) \le 0,$$

thus

$$\Delta y_j \sum_{\beta=1}^{k+1} \hat{\omega}_\beta a(x_{i+\frac{1}{2}}, \hat{y}_\beta) u^+_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) + \Delta x_i \sum_{\alpha=1}^{k+1} \hat{\omega}_\alpha b(\hat{x}_\alpha, y_{j+\frac{1}{2}}) u^+_{i,j}(\hat{x}_\alpha, y^-_{j+\frac{1}{2}}) \leq$$
$$\Delta y_j \sum_{\beta=1}^{k+1} \hat{\omega}_\beta a(x_{i+\frac{1}{2}}, \hat{y}_\beta) u^-_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) + \Delta x_i \sum_{\alpha=1}^{k+1} \hat{\omega}_\alpha b(\hat{x}_\alpha, y_{j+\frac{1}{2}}) u^-_{i,j}(\hat{x}_\alpha, y^-_{j+\frac{1}{2}}),$$

which implies

$$\sum_{\beta=1}^{k+1} \hat{\omega}_\beta u^+_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) + \sum_{\alpha=1}^{k+1} \hat{\omega}_\alpha u^+_{i,j}(\hat{x}_\alpha, y^-_{j+\frac{1}{2}}) \leq \frac{\rho \max\{a^*, b^*\}}{\min\{a_*, b_*\}} \left(\sum_{\beta=1}^{k+1} \hat{\omega}_\beta u^-_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) + \sum_{\alpha=1}^{k+1} \hat{\omega}_\alpha u^-_{i,j}(\hat{x}_\alpha, y^-_{j+\frac{1}{2}}) \right) = O(h^{k+1}).$$

By the definition of $\hat{u}^b_{i,j}$, we have $||\hat{u}^b_{i,j}||_{L^\infty(K_{i,j})} = O(h^{k+1})$, therefore $||e_{1,3}||_{L^\infty(K_{i,j})} = O(h^{k+1})$.

Case II: $\theta^1_{i,j} > 0$. We have $LHS^b(u^b_{i,j}) > 0$ and $\theta^1_{i,j} = \frac{LHS^b(u^b_{i,j})}{LHS^b(\hat{u}^b_{i,j})}$. Therefore,

$$\begin{aligned}
|e_{1,3}| =& (1-\theta^1_{i,j})|\hat{u}^b_{i,j}| \\
=& \frac{LHS^b(\hat{u}^b_{i,j} - u^b_{i,j})}{LHS^b(\hat{u}^b_{i,j})} |\hat{u}^b_{i,j}| \\
=& \left(\Delta y_j \sum_{\beta=1}^{k+1} \hat{\omega}_\beta a(x_{i+\frac{1}{2}}, \hat{y}_\beta) \hat{u}^b_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) + \Delta x_i \sum_{\alpha=1}^{k+1} \hat{\omega}_\alpha b(\hat{x}_\alpha, y_{j+\frac{1}{2}}) \hat{u}^b_{i,j}(\hat{x}_\alpha, y^-_{j+\frac{1}{2}}) \right)^{-1} \times \\
& \left(\Delta y_j \sum_{\beta=1}^{k+1} \hat{\omega}_\beta a(x_{i+\frac{1}{2}}, \hat{y}_\beta) \left(\hat{u}^b_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) - u^b_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) \right) \right. \\
& \left. + \Delta x_i \sum_{\alpha=1}^{k+1} \hat{\omega}_\alpha b(\hat{x}_\alpha, y_{j+\frac{1}{2}}) \left(\hat{u}^b_{i,j}(\hat{x}_\alpha, y^-_{j+\frac{1}{2}}) - u^b_{i,j}(\hat{x}_\alpha, y^-_{j+\frac{1}{2}}) \right) \right) \cdot |\hat{u}^b_{i,j}| \\
\leq& \frac{\left(\Delta y_j \sum_{\beta=1}^{k+1} \hat{\omega}_\beta a(x_{i+\frac{1}{2}}, \hat{y}_\beta) + \Delta x_i \sum_{\alpha=1}^{k+1} \hat{\omega}_\alpha b(\hat{x}_\alpha, y_{j+\frac{1}{2}}) \right) ||e_{1,2}||_{L^\infty(K_{i,j})}}{\Delta y_j \sum_{\beta=1}^{k+1} \hat{\omega}_\beta a(x_{i+\frac{1}{2}}, \hat{y}_\beta) \hat{u}^b_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) + \Delta x_i \sum_{\alpha=1}^{k+1} \hat{\omega}_\alpha b(\hat{x}_\alpha, y_{j+\frac{1}{2}}) \hat{u}^b_{i,j}(\hat{x}_\alpha, y^-_{j+\frac{1}{2}})} \cdot |\hat{u}^b_{i,j}|
\end{aligned} \tag{3.7}$$

In particular, $e_{1,3}(\hat{x}_{\gamma_1}, \hat{y}_{\gamma_2}) = 0$ for $\gamma_1, \gamma_2 = 1, 2, \ldots, k$, since $\hat{u}^b_{i,j} = 0$ at these points by definition. Moreover, for $\gamma = 1, 2, \ldots, k+1$, we have the following estimates from (3.7),

$$\begin{aligned}
|e_{1,3}(x^-_{i+\frac{1}{2}}, \hat{y}_\gamma)| \leq& \frac{\left(\Delta y_j \sum_{\beta=1}^{k+1} \hat{\omega}_\beta a(x_{i+\frac{1}{2}}, \hat{y}_\beta) + \Delta x_i \sum_{\alpha=1}^{k+1} \hat{\omega}_\alpha b(\hat{x}_\alpha, y_{j+\frac{1}{2}}) \right) ||e_{1,2}||_{L^\infty(K_{i,j})}}{\Delta y_j \sum_{\beta=1}^{k+1} \hat{\omega}_\beta a(x_{i+\frac{1}{2}}, \hat{y}_\beta) \hat{u}^b_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\beta) + \Delta x_i \sum_{\alpha=1}^{k+1} \hat{\omega}_\alpha b(\hat{x}_\alpha, y_{j+\frac{1}{2}}) \hat{u}^b_{i,j}(\hat{x}_\alpha, y^-_{j+\frac{1}{2}})} \hat{u}^b_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\gamma) \\
\leq& \frac{(\Delta y_j a^* + \Delta x_i b^*) ||e_{1,2}||_{L^\infty(K_{i,j})}}{\Delta y_j a_* \sum_{\beta=1}^{k+1} \hat{\omega}_\beta \hat{u}^b_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\beta)} \hat{u}^b_{i,j}(x^-_{i+\frac{1}{2}}, \hat{y}_\gamma) \\
\leq& \frac{\rho\,(a^* + b^*)}{a_* \hat{\omega}_\gamma} ||e_{1,2}||_{L^\infty(K_{i,j})} = O(h^{k+1}),
\end{aligned} \tag{3.8}$$

and similarly, $|e_{1,3}(\hat{x}_\gamma, y^-_{j+\frac{1}{2}})| = O(h^{k+1}), \gamma = 1, 2, \ldots, k+1$. Therefore, following the similar argument as (2.9), we have $||e_{1,3}||_{L^\infty(K_{i,j})} = \Lambda^2_k \cdot O(h^{k+1}) = O(h^{k+1})$.

To sum up, we have $||e_1||_{L^\infty(K_{i,j})} \le ||e_{1,1}||_{L^\infty(K_{i,j})} + ||e_{1,2}||_{L^\infty(K_{i,j})} + ||e_{1,3}||_{L^\infty(K_{i,j})} = O(h^{k+1})$.

We now estimate e_2 as follows. If $\lambda = 0$, then $\theta_{i,j}^2 = 1$, thus $e_2 = (1-\theta_{i,j}^2)\hat{u}_{i,j} = 0$. If $\lambda > 0$, we have

$$
\begin{aligned}
&|e_2(x,y)| = \left(1 - \frac{LHS(u_{i,j})}{LHS(\hat{u}_{i,j})}\right)|\hat{u}_{i,j}(x,y)| = \frac{LHS(\hat{u}_{i,j}-u_{i,j})}{LHS(\hat{u}_{i,j})}|\hat{u}_{i,j}(x,y)| \\
&= \left(\lambda\Delta x_i\Delta y_j \sum_{\alpha=1}^{k+1}\sum_{\beta=1}^{k+1}\hat{\omega}_\alpha\hat{\omega}_\beta\left(\hat{u}_{i,j}(\hat{x}_\alpha,\hat{y}_\beta) - u_{i,j}(\hat{x}_\alpha,\hat{y}_\beta)\right) + \Delta y_j\sum_{\beta=1}^{k+1}\hat{\omega}_\beta a(x_{i+\frac{1}{2}},\hat{y}_\beta)\left(\hat{u}_{i,j}(x^-_{i+\frac{1}{2}},\hat{y}_\beta) - u_{i,j}(x^-_{i+\frac{1}{2}},\hat{y}_\beta)\right)\right. \\
&\left.+\Delta x_i\sum_{\alpha=1}^{k+1}\hat{\omega}_\alpha b(\hat{x}_\alpha,y_{j+\frac{1}{2}})\left(\hat{u}_{i,j}(\hat{x}_\alpha,y^-_{j+\frac{1}{2}}) - u_{i,j}(\hat{x}_\alpha,y^-_{j+\frac{1}{2}})\right)\right)\cdot\left(\lambda\Delta x_i\Delta y_j\sum_{\alpha=1}^{k+1}\sum_{\beta=1}^{k+1}\hat{\omega}_\alpha\hat{\omega}_\beta\hat{u}_{i,j}(\hat{x}_\alpha,\hat{y}_\beta)\right. \\
&\left.+\ \Delta y_j\sum_{\beta=1}^{k+1}\hat{\omega}_\beta a(x_{i+\frac{1}{2}},\hat{y}_\beta)\hat{u}_{i,j}(x^-_{i+\frac{1}{2}},\hat{y}_\beta) + \Delta x_i\sum_{\alpha=1}^{k+1}\hat{\omega}_\alpha b(\hat{x}_\alpha,y_{j+\frac{1}{2}})\hat{u}_{i,j}(\hat{x}_\alpha,y^-_{j+\frac{1}{2}})\right)^{-1}\cdot|\hat{u}_{i,j}(x,y)| \\
&\le (\lambda\Delta x_i\Delta y_j + a^*\Delta y_j + b^*\Delta x_i)\cdot||e_1||_{L^\infty(K_{i,j})}\cdot\left(\lambda\Delta x_i\Delta y_j\sum_{\alpha=1}^{k+1}\sum_{\beta=1}^{k+1}\hat{\omega}_\alpha\hat{\omega}_\beta\hat{u}_{i,j}(\hat{x}_\alpha,\hat{y}_\beta)\right. \\
&\left.+a_*\Delta y_j\sum_{\beta=1}^{k+1}\hat{\omega}_\beta\hat{u}_{i,j}(x^-_{i+\frac{1}{2}},\hat{y}_\beta) + b_*\Delta x_i\sum_{\alpha=1}^{k+1}\hat{\omega}_\alpha\hat{u}_{i,j}(\hat{x}_\alpha,y^-_{j+\frac{1}{2}})\right)^{-1}\cdot|\hat{u}_{i,j}(x,y)|
\end{aligned}
\tag{3.9}
$$

In particular, at the Gauss-Radau points, we have the following estimates from (3.9),

$$
\begin{aligned}
|e_2(\hat{x}_{\gamma_1},\hat{y}_{\gamma_2})| &\le (\lambda\Delta x_i\Delta y_j + a^*\Delta y_j + b^*\Delta x_i)\cdot||e_1||_{L^\infty(K_{i,j})}\cdot\left(\lambda\Delta x_i\Delta y_j\sum_{\alpha=1}^{k+1}\sum_{\beta=1}^{k+1}\hat{\omega}_\alpha\hat{\omega}_\beta\hat{u}_{i,j}(\hat{x}_\alpha,\hat{y}_\beta)\right)^{-1}\cdot\hat{u}_{i,j}(\hat{x}_{\gamma_1},\hat{y}_{\gamma_2}) \\
&\le \frac{\lambda\rho^2h^2 + a^*\rho h + b^*\rho h}{\lambda h^2\hat{\omega}_{\gamma_1}\hat{\omega}_{\gamma_2}}||e_1||_{L^\infty(K_{i,j})} = \left(\rho^2\hat{\omega}_{\gamma_1}^{-1}\hat{\omega}_{\gamma_2}^{-1} + \frac{a^*\rho + b^*\rho}{\lambda\hat{\omega}_{\gamma_1}\hat{\omega}_{\gamma_2}}\frac{1}{h}\right)||e_1||_{L^\infty(K_{i,j})} \\
&= O(h^k), \quad \text{for } \gamma_1,\gamma_2 = 1,2,\ldots,k+1,
\end{aligned}
\tag{3.10}
$$

therefore, following the similar argument as (2.9), we have $||e_2||_{L^\infty(K_{i,j})} = \Lambda_k^2\cdot O(h^k) = O(h^k)$.

In particular, on the downstream edge $I_{i+\frac{1}{2}}$, it follows from (3.9) that

$$
\begin{aligned}
&|e_2(x_{i+\frac{1}{2}},\hat{y}_\gamma)| \le (\lambda\Delta x_i\Delta y_j + a^*\Delta y_j + b^*\Delta x_i)\cdot||e_1||_{L^\infty(K_{i,j})}\cdot\left(a_*\Delta y_j\sum_{\beta=1}^{k+1}\hat{\omega}_\beta\hat{u}_{i,j}(x^-_{i+\frac{1}{2}},\hat{y}_\beta)\right)^{-1}\cdot\hat{u}_{i,j}(x^-_{i+\frac{1}{2}},\hat{y}_\gamma) \\
&\le (\lambda\Delta x_i\Delta y_j + a^*\Delta y_j + b^*\Delta x_i)\cdot||e_1||_{L^\infty(K_{i,j})}\cdot\left(a_*\Delta y_j\hat{\omega}_\gamma\hat{u}_{i,j}(x^-_{i+\frac{1}{2}},\hat{y}_\gamma)\right)^{-1}\cdot\hat{u}_{i,j}(x^-_{i+\frac{1}{2}},\hat{y}_\gamma) \\
&\le \frac{\lambda\rho^2h + a^*\rho + b^*\rho}{a_*\hat{\omega}_\gamma}||e_1||_{L^\infty(K_{i,j})} = O(h^{k+1}),
\end{aligned}
\tag{3.11}
$$

for $\gamma = 1, 2, \ldots, k+1$. Similarly, on the downstream edge $J_{j+\frac{1}{2}}$, we have $|e_2(\hat{x}_\gamma, y_{j+\frac{1}{2}})| = O(h^{k+1})$, for $\gamma = 1, 2, \ldots, k+1$. Following the same lines as in (2.9), we have the estimate $||e_2||_{L^\infty(I_{i+\frac{1}{2}} \cup J_{j+\frac{1}{2}})} = \Lambda_k \cdot O(h^{k+1}) = O(h^{k+1})$. Thus $||\tilde{u}_{i,j} - u_{i,j}||_{L^\infty(I_{i+\frac{1}{2}} \cup J_{j+\frac{1}{2}})} = O(h^{k+1})$ by the triangle inequality.

Gathering all results above, we finish the proof of Lemma 3.1. □

The type-2 limiter is defined as follows,

$$\tilde{u}_{i,j}(x,y) = \theta_{i,j}\hat{u}_{i,j}(x,y), \quad \hat{u}_{i,j}(x,y) = u_{i,j}(x,y) + \epsilon_{i,j}, \tag{3.12}$$

where $\epsilon_{i,j} = -\min\{\min_{(x,y)\in S} u_{i,j}(x,y), 0\}$, $S \subset K_{i,j}$ is the set of points where we want to preserve the positivity of solutions, and $\theta_{i,j} = \frac{LHS(u_{i,j})}{LHS(\hat{u}_{i,j})} \in [0,1]$.

We have the accuracy result for the conservative positivity-preserving limiter as follows:

Lemma 3.2. *Consider the solution $u_{i,j}$ of the scheme (3.2) with an L^∞-error of $O(h^{k+1})$. If $\lambda = 0$, the error introduced by the limiter (3.12) is $||\tilde{u}_{i,j} - u_{i,j}||_{L^\infty(K_{i,j})} = O(h)$. If $\lambda > 0$, the error introduced by the limiter (3.12) is $||\tilde{u}_{i,j} - u_{i,j}||_{L^\infty(K_{i,j})} = O(h^k)$. Nevertheless, on the downstream edges, the errors in both cases are optimal, i.e. $||\tilde{u}_{i,j} - u_{i,j}||_{L^\infty(I_{i+\frac{1}{2}} \cup J_{j+\frac{1}{2}})} = O(h^{k+1})$.*

Proof. For simplicity, we assume $S = K_{i,j}$. Same as (3.5), we decompose the error as $e = u_{i,j} - \tilde{u}_{i,j} = (u_{i,j} - \hat{u}_{i,j}) + (\hat{u}_{i,j} - \tilde{u}_{i,j}) = e_1 + e_2$. It is clear that $|e_1| = \epsilon_{i,j} = O(h^{k+1})$. Consider $e_2 = (1 - \theta_{i,j})\,\hat{u}_{i,j}$. If $\lambda = 0$, we have $||e_2||_{L^\infty(K_{i,j})} = (1 - \theta_{i,j})\,||\hat{u}_{i,j}||_{L^\infty(K_{i,j})} = O(h)$, since $||\hat{u}_{i,j}||_{L^\infty(K_{i,j})} = O(h)$ if $\theta_{i,j} < 1$. If $\lambda > 0$, we have the same estimates for e_2 as (3.9) and (3.10). The estimates for e_2 on the downstream edges are exactly the same as (3.11) for both the cases $\lambda = 0$ and $\lambda > 0$.

Collecting all results above, we finish the proof of Lemma 3.2. □

Since the accuracy of both type-1 and type-2 limiters is optimal on the downstream edges, we do not need to worry about the pollution of the non-optimal errors introduced by the limiters to the downstream cells. Thus we have the following theorem for the positivity-preserving DG method of the equation (3.1).

Theorem 3.3. *For the linear stationary hyperbolic equation (3.1), if the source term and inflow boundary conditions are nonnegative, then the solution of the scheme (3.2) modified by the limiter (3.4) or (3.12) is nonnegative, with the local accuracy established in Lemmas 3.1 and 3.2, respectively.*

Remark 3.1. *In particular, in the space-time DG discretization for the equation of the form $u_t + (a(x)u)_x = s(x,t)$, the accuracy of the solution at the terminal time is optimal, as the terminal time is indeed an outflow boundary.*

4 Linear stationary hyperbolic equations in two dimensions on triangular meshes

In this section, we study the high order conservative positivity-preserving discontinuous Galerkin method in two space dimensions on triangular meshes for the linear stationary hyperbolic equation (3.1) with nonnegative source term and the inflow boundary condition $u|_{\Gamma^{\text{in}}}(x,y) = g(x,y) \geq 0$, where $\Gamma^{\text{in}} \subset \partial\Omega$ is the inflow boundary. We still assume $\lambda \geq 0$ in (3.1) but $a(x,y)$ and $b(x,y)$ are not necessarily positive (or negative).

Consider a regular triangulation Ω_h of Ω which satisfies $\text{diam}(K) \leq \rho h, \forall K \in \Omega_h$ for some $\rho \geq 1$ independent of the refinement, where $\text{diam}(K)$ is the diameter of an element K, $h = \min_{K\in\Omega_h} h_K$ and h_K is the radius of the largest ball inscribed in K. For any triangle element $K \in \Omega_h$, we denote by $|K|$ the area of K, and $e_K^i, i = 1,2,3$ the three edges of K, with length ℓ_K^i, unit outer normal $n_K^i = (n_{x,K}^i, n_{y,K}^i)^T$ and neighboring cells K_i, $i = 1,2,3$. We assume that the coefficients $a(x,y)$ and $b(x,y)$ in (3.1) satisfy $c_* \leq \left|a(x,y)n_{x,K}^i + b(x,y)n_{y,K}^i\right| \leq c^*, \forall K \in \Omega_h, (x,y) \in \Omega, i = 1,2,3$, for some positive constants c_*, c^*. This assumption was adopted in the optimal order error estimate for the DG method in [26], as the optimal accuracy is unavailable for general meshes [23]. The assumption can be satisfied, for instance, by the conditions on the coefficients $a(x,y), b(x,y)$ in Section 3, together with the triangulation obtained by splitting each cell therein from the skew diagonal of cells, see Figure 2 for an illustration. The function space V of the P^k-DG scheme is defined as

$$V = \left\{v \in L^2(\Omega) : v|_K \in P^k(K), \forall K \in \Omega_h\right\},$$

where $P^k(K)$ denotes the space of polynomials of order no greater than k on the element K. We define the cell average of $v \in V$ on K as $\bar{v}_K = \frac{1}{|K|}\iint_K v(x,y)dxdy$, and denote by $v_K = v|_K$ for $v \in V$.

To save space, we only discuss the case that e_K^1 is the upstream edge and e_K^2, e_K^3 are the downstream edges, as the discussion of the case of two upstream edges and one downstream edge is almost the same with the first case.

The positivity-preserving P^k-DG scheme of the equation (3.1) on triangular meshes is to find $u \in V$, such that

$$\begin{aligned} &-\iint_K \left(auv_x + buv_y - \lambda uv\right) dxdy + \int_{e_K^2} \left(an_{x,K}^2 + bn_{y,K}^2\right) u_K v ds + \int_{e_K^3} \left(an_{x,K}^3 + bn_{y,K}^3\right) u_K v ds \\ &= -\int_{e_K^1} \left(an_{x,K}^1 + bn_{y,K}^1\right) \tilde{u}_{K_1} v ds + \iint_K svdxdy, \quad \forall v \in P^k(K), \end{aligned} \tag{4.1}$$

for $K \in \Omega_h$, where we define $\tilde{u}_{K_1}|_{e_K^1} = \mathcal{I}(g)$ if $e_K^1 \subset \Gamma^{\text{in}}$, with $\mathcal{I}$ denoting the polynomial interpolation at the quadrature points on cell interfaces. In the computation, we solve u_K on cell K based on the modified solution on upstream cells. Once u_K is obtained, we employ the positivity-preserving limiter to obtain the modified solution $\tilde{u}_K$, and use it in the computation on the downstream cells.

If we take $v = 1$ on K in the scheme (4.1), the following equation satisfied by the local mass can be obtained

$$\lambda|K|\bar{u}_K + \int_{e_K^2} \left(an_{x,K}^2 + bn_{y,K}^2\right) u_K ds + \int_{e_K^3} \left(an_{x,K}^3 + bn_{y,K}^3\right) u_K ds = -\int_{e_K^1} \left(an_{x,K}^1 + bn_{y,K}^1\right) \tilde{u}_{K_1} ds + |K|\bar{s}_K. \tag{4.2}$$

We define $LHS(w_K) = \lambda|K|\bar{w}_K + \int_{e_K^2} \left(an_{x,K}^2 + bn_{y,K}^2\right) w_K ds + \int_{e_K^3} \left(an_{x,K}^3 + bn_{y,K}^3\right) w_K ds$, for $w_K \in P^k(K)$, to be the amount of local mass of w_K on K. Since $\tilde{u}_{K_1} \geq 0$ and $an_{x,K}^1 + bn_{y,K}^1 < 0$ on the upstream edge in (4.2), we have the $LHS(u_K) \geq 0$.

Due to the lack of suitable quadrature rules, we do not have the type-1 limiter available. The type-2 limiter is defined as follows,

$$\tilde{u}_K(x,y) = \theta_K \hat{u}_K(x,y), \quad \hat{u}_K(x,y) = u_K(x,y) + \epsilon_K, \tag{4.3}$$

where $\epsilon_K = -\min\{\min_{(x,y)\in S} u_K(x,y), 0\}$, $S \subset K$ is the set of points where we want to preserve the positivity of solutions, and $\theta_K = \frac{LHS(u_K)}{LHS(\hat{u}_K)} \in [0,1]$.

We have the accuracy result for the conservative positivity-preserving limiter as follows:

Lemma 4.1. *Consider the solution u_K of the scheme (4.1) with an L^∞-error of $O(h^{k+1})$. If $\lambda = 0$, the error introduced by the limiter (4.3) is $||\tilde{u}_K - u_K||_{L^\infty(K)} = O(h)$. If $\lambda > 0$, the error introduced by the limiter (4.3) is $||\tilde{u}_K - u_K||_{L^\infty(K)} = O(h^k)$. Nevertheless, on the downstream edges, the errors in both cases are optimal, i.e. $||\tilde{u}_K - u_K||_{L^\infty(e_K^2 \cup e_K^3)} = O(h^{k+1})$.*

Proof. For simplicity, we assume $S = K$. We decompose the error as

$$e = u_K - \tilde{u}_K = (u_K - \hat{u}_K) + (\hat{u}_K - \tilde{u}_K) = e_1 + e_2 \tag{4.4}$$

It is clear that $||e_1||_{L^\infty(K)} = \epsilon_K = O(h^{k+1})$.

For e_2, we have $e_2 = \hat{u}_K - \tilde{u}_K = (1-\theta_K)\hat{u}_K$. If $\lambda = 0$, we have $||e_2||_{L^\infty(K)} = (1-\theta_K)||\hat{u}_K||_{L^\infty(K)} = O(h)$,

since $||\hat{u}_K||_{L^\infty(K)} = O(h)$ if $\theta_K < 1$. If $\lambda > 0$, we have the estimate for e_2 as follows,

$$\begin{aligned}
|e_2| &= (1-\theta_K)\hat{u}_K = \left(1 - \frac{LHS(u_K)}{LHS(\hat{u}_K)}\right)\hat{u}_K = \frac{LHS(\hat{u}_K - u_K)}{LHS(\hat{u}_K)}\hat{u}_K \\
&= \frac{\lambda|K|\left(\bar{\hat{u}}_K - \bar{u}_K\right) + \int_{e_K^2}\left(an_{x,K}^2 + bn_{y,K}^2\right)(\hat{u}_K - u_K)\,ds + \int_{e_K^3}\left(an_{x,K}^3 + bn_{y,K}^3\right)(\hat{u}_K - u_K)\,ds}{\lambda|K|\bar{\hat{u}}_K + \int_{e_K^2}\left(an_{x,K}^2 + bn_{y,K}^2\right)\hat{u}_K ds + \int_{e_K^3}\left(an_{x,K}^3 + bn_{y,K}^3\right)\hat{u}_K ds}\hat{u}_K \\
&\le \frac{\left(\lambda|K| + c^*\ell_K^2 + c^*\ell_K^3\right)||e_1||_{L^\infty(K)}}{\lambda|K|\bar{\hat{u}}_K + c_*\ell_K^2\bar{\hat{u}}_{e_K^2} + c_*\ell_K^3\bar{\hat{u}}_{e_K^3}}\hat{u}_K,
\end{aligned} \tag{4.5}$$

where $\bar{v}_{e_K^i} = \frac{1}{\ell_K^i}\int_{e_K^i} v_K ds$, for $v \in V$, $i = 1, 2, 3$.

By the equivalence of norms in the finite-dimensional space $P^k(K)$ and the rescaling argument, we have $||v||_{L^\infty(K)} \le \frac{C_k}{|K|}||v||_{L^1(K)}$ and $||v||_{L^\infty(e_K^i)} \le \frac{C_k'}{\ell_K^i}||v||_{L^1(e_K^i)}$, $\forall v \in P^k(K), i = 1, 2, 3$, for some positive constants C_k and C_k' depending only on k. Therefore,

$$\begin{aligned}
||e_2||_{L^\infty(K)} &\le \frac{\left(\lambda|K| + c^*\ell_K^2 + c^*\ell_K^3\right)||e_1||_{L^\infty(K)}}{\lambda|K|}\frac{||\hat{u}_K||_{L^\infty(K)}}{\bar{\hat{u}}_K} \\
&\le \frac{\left(\lambda|K| + 2c^*\rho h\right)||e_1||_{L^\infty(K)}}{\lambda|K|}C_k \\
&\le C_k\left(1 + \frac{2c^*\rho}{\pi\lambda}\frac{1}{h}\right)||e_1||_{L^\infty(K)} \\
&= O(h^k),
\end{aligned} \tag{4.6}$$

where we have used the fact that $\hat{u}_K \ge 0$. Moreover, we have

$$\begin{aligned}
||e_2||_{L^\infty(e_K^2)} &\le \frac{\left(\lambda|K| + c^*\ell_K^2 + c^*\ell_K^3\right)||e_1||_{L^\infty(K)}}{c_*\ell_K^2}\frac{||\hat{u}_K||_{L^\infty(e_K^2)}}{\bar{\hat{u}}_{e_K^2}} \\
&\le \frac{\left(\lambda\rho^2h^2 + 2c^*\rho h\right)||e_1||_{L^\infty(K)}}{c_* h}C_k' \\
&= C_k'\frac{\left(\lambda\rho^2 h + 2c^*\rho\right)}{c_*}||e_1||_{L^\infty(K)} \\
&= O(h^{k+1}),
\end{aligned} \tag{4.7}$$

and, similarly, $||e_2||_{L^\infty(e_K^3)} = O(h^{k+1})$.

Gathering all results above and using triangle inequalities, we finish the proof of Lemma 4.1. □

Since the accuracy of the limiter (4.3) is optimal on the downstream edges, we have the following theorem for the positivity-preserving DG method of the equation (3.1).

Theorem 4.2. *For the linear stationary hyperbolic equation* (3.1)*, if the source term and inflow boundary condition are nonnegative, then the solution of the scheme* (4.1) *modified by the limiter* (4.3) *is nonnegative, with the local accuracy established in Lemma 4.1.*

5 Nonlinear stationary hyperbolic equations with invariant sign of $f'(u)$ in one dimension

In this section, we study the high order conservative positivity-preserving discontinuous Galerkin method for the nonlinear stationary hyperbolic equation

$$f(u)_x + \lambda u = s(x), \quad x \in \Omega = (0,1), \tag{5.1}$$

where $0 \le f'(u) \le a^*, \forall u$, $f'(u)$ has at most countably many zeros, and $\lambda, s(x) \ge 0$. We assign the inflow boundary condition $u(0) = u_0 \ge 0$ for the equation. We would like to note that, the assumption on invariant sign of $f'(u)$ for all u is essential, otherwise the stationary hyperbolic equation may need boundary conditions from both sides for the problem to be well-posed, see [17, 25] for instance. This condition is also necessary for the limiters to be well-defined.

We adopt the partition for Ω and the function space V exactly the same as in Section 2, as well as the notations if not otherwise stated.

The positivity-preserving P^k-DG scheme of the equation (5.1) is to find $u \in V$, such that

$$-\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} f(u)v_x dx + f(u^-_{j+\frac{1}{2}})v^-_{j+\frac{1}{2}} + \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \lambda uv dx = f(\tilde{u}^-_{j-\frac{1}{2}})v^+_{j-\frac{1}{2}} + \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} sv dx, \quad \forall v \in P^k(I_j), \tag{5.2}$$

for $j = 1, 2, \ldots, N$, where we define $u^-_{\frac{1}{2}} = u_0$. Note that the upstream cells adopt the modified solution in the scheme.

If we take the test function $v = 1$ on I_j in the scheme (5.2), the following equation satisfied by the local mass is obtained,

$$\lambda \Delta x_j \bar{u}_j + f(u^-_{j+\frac{1}{2}}) = f(\tilde{u}^-_{j-\frac{1}{2}}) + \Delta x_j \bar{s}_j. \tag{5.3}$$

Same as the linear case, we define $LHS(w_j) = \lambda \Delta x_j \bar{w}_j + f(w^-_{j+\frac{1}{2}})$, for $w_j \in P^k(I_j)$, to be the amount of local mass of w_j on I_j. A notable difference is that, we no longer have $LHS(u_j) \ge 0$.

The type-1 limiter for u_j is defined as follows,

$$\tilde{u}_j(x) = \theta_j \hat{u}_j(x), \quad \hat{u}_j(x) = \sum_{\alpha=1}^{k+1} u_j^+(\hat{x}_\alpha)\ell_\alpha(x), \tag{5.4}$$

where $\theta_j \in [0,1]$ is taken such that the local mass is conservative, i.e. $LHS(\tilde{u}_j) = LHS(u_j)$. Same as before, the type-1 limiter must be used in cooperation with the Gauss-Radau quadrature.

If $\lambda > 0$, $\theta_j \in [0,1]$ is uniquely determined. To see this, we define $h(\theta) = LHS(\theta\hat{u}_j) - LHS(u_j)$. It is clear that $h(0) = f(0) - f(\tilde{u}^-_{j-\frac{1}{2}}) - \Delta x_j \bar{s}_j \le 0$, $h(1) = \lambda\Delta x_j \sum_{\alpha=1}^{k+1} \hat{\omega}_\alpha u_j^-(\hat{x}_\alpha) + f((u^-_{j+\frac{1}{2}})^+) - f(u^-_{j+\frac{1}{2}}) \ge 0$, and $h'(\theta) > 0$ for $\theta \in [0,1]$. Therefore, the existence and uniqueness of θ_j is guaranteed by the mean value theorem and monotonicity of $h(\theta)$.

If $\lambda = 0$, the identity (5.3) becomes $f(u^-_{j+\frac{1}{2}}) = f(\tilde{u}^-_{j-\frac{1}{2}}) + \Delta x_j \bar{s}_j$, which implies $f(u^-_{j+\frac{1}{2}}) \ge f(\tilde{u}^-_{j-\frac{1}{2}})$ since $\bar{s}_j \ge 0$. From the monotonicity of f, we have $u^-_{j+\frac{1}{2}} \ge \tilde{u}^-_{j-\frac{1}{2}} \ge 0$. We always take $\theta_j = 1$ as $u^-_{j+\frac{1}{2}} \ge 0$, which implies $LHS(\hat{u}_j) = f(u^-_{j+\frac{1}{2}}) = LHS(u_j)$.

Moreover, we have the accuracy result of the limiter as follows:

Lemma 5.1. *Consider the solution u_j of the scheme (5.2) with an L^∞-error of $O(\Delta x_j^{k+1})$. If $\lambda = 0$, the error introduced by the limiter (5.4) is $||\tilde{u}_j - u_j||_{L^\infty(I_j)} = O(\Delta x_j^{k+1})$. If $\lambda > 0$, the error introduced by the limiter (5.4) is $||\tilde{u}_j - u_j||_{L^\infty(I_j)} = O(\Delta x_j^k)$, but the error is optimal at the downstream point, i.e. $|\tilde{u}^-_{j+\frac{1}{2}} - u^-_{j+\frac{1}{2}}| = O(\Delta x_j^{k+1})$.*

Proof. From $\theta_j\lambda\Delta x_j\bar{\hat{u}}_j + f(\theta_j\hat{u}^-_{j+\frac{1}{2}}) = \lambda\Delta x_j\bar{u}_j + f(u^-_{j+\frac{1}{2}})$, we have the expression of θ_j as follows,

$$\theta_j = \frac{\lambda\Delta x_j\bar{u}_j + f'(c_{j+\frac{1}{2}})u^-_{j+\frac{1}{2}}}{\lambda\Delta x_j\bar{\hat{u}}_j + f'(c_{j+\frac{1}{2}})\hat{u}^-_{j+\frac{1}{2}}}, \tag{5.5}$$

where $c_{j+\frac{1}{2}} \in [u^-_{j+\frac{1}{2}}, \hat{u}^-_{j+\frac{1}{2}}]$ satisfies the Lagrange mean value theorem $f(\theta_j\hat{u}^-_{j+\frac{1}{2}}) - f(u^-_{j+\frac{1}{2}}) = f'(c_{j+\frac{1}{2}})(\theta_j\hat{u}^-_{j+\frac{1}{2}} - u^-_{j+\frac{1}{2}})$. Then the estimates are almost the same to those in the proof of Lemma 2.1, except that $a(x_{j+\frac{1}{2}})$ is replaced by $f'(c_{j+\frac{1}{2}})$. □

The type-2 limiter for u_j is defined as follows,

$$\tilde{u}_j(x) = \theta_j\hat{u}_j(x), \quad \hat{u}_j(x) = u_j(x) + \epsilon_j, \tag{5.6}$$

where $\epsilon_j = -\min\{\min_{x\in S} u_j(x), 0\}$, $S \subset I_j$ is the set of points where we want to preserve the positivity of solutions, and $\theta_j \in [0,1]$ is uniquely determined by $LHS(\tilde{u}_j) = LHS(u_j)$.

We have the accuracy result for the limiter as follows,

Lemma 5.2. *Consider the solution u_j of the scheme (5.2) with an L^∞-error of $O(\Delta x_j^{k+1})$. If $\lambda = 0$, the error introduced by the limiter (5.6) is $||\tilde{u}_j - u_j||_{L^\infty(I_j)} = O(\Delta x_j)$. If $\lambda > 0$, the error introduced by the limiter (5.6) is $||\tilde{u}_j - u_j||_{L^\infty(I_j)} = O(\Delta x_j^k)$. Nevertheless, at the downstream point, the errors in both cases are optimal, i.e. $|\tilde{u}^-_{j+\frac{1}{2}} - u^-_{j+\frac{1}{2}}| = O(\Delta x_j^{k+1})$.*

Proof. We have the same expression of θ_j as in (5.5). Therefore the estimates are almost the same with those in the proof of Lemma 2.2, except that $a(x_{j+\frac{1}{2}})$ is replaced by $f'(c_{j+\frac{1}{2}})$. □

Since the accuracy of both type-1 and type-2 limiters is optimal at the downstream point of cells, the assumption on the optimal accuracy of the unmodulated DG solution is appropriate. Collecting the Lemma 5.1 and 5.2, we attain the following theorem for the positivity-preserving DG method for the equation (5.1).

Theorem 5.3. *For the nonlinear stationary hyperbolic equation* (5.1)*, if the source term and inflow boundary condition are nonnegative, then the solution of the scheme* (5.2) *modified by the limiters* (5.4) *or* (5.6) *is nonnegative, with the local accuracy established in Lemma 5.1 and 5.2, respectively.*

6 Numerical tests

In this section, we show the accuracy and effectiveness of the conservative positivity-preserving DG methods established in previous sections for stationary hyperbolic equations and time-dependent problems with implicit time discretization by ample numerical tests. Most of the tests are taken from [21, 30, 33]. For simplicity, the triangular meshes adopted in the two dimensional tests are obtained by splitting the rectangular grids by the diagonals of every cells, see Figure 2 for an illustration of a 6×6 mesh.

To save space, we only present the results of the type-2 limiters, as those of the type-1 limiters are almost the same (even thought the type-1 limiter is formally more accurate than the type-2). For the P^k-DG scheme in one dimension, we adopt the $(k+1)$-point Gauss-Legendre rule for volume integration. For the Q^k-DG scheme on rectangular meshes, we use the tensor product of $(k+1)$-point Gauss-Legendre rules for volume integration and $(k+1)$-point Gauss-Legendre rule for edge integration. For the P^k-DG scheme on triangular meshes, we adopt the quadrature rule developed in [28] with 36 points, which is exact for polynomials of degree up to 13 for volume integration, and $(k+1)$-point Gauss-Legendre rule for edge integration.

In one dimension, we compute the errors in L^1 norm using the composite midpoint rule with 100 uniformly distributed points in each cell. The errors in L^∞ norm are computed by evaluating the maximum norm at these points. Similarly, on rectangular meshes, we compute the errors in L^1 norm using the composite midpoint rule with 20 uniformly distributed points in each dimension of a cell. The errors in L^∞ norm are computed by evaluating the maximum norm of errors at these points. On triangular meshes, we compute the L^1 norm using the same quadrature rule used for the volume integration of the scheme. The L^∞ norm

is computed by evaluating the maximum norm at these points.

We would like to note that, though the sub-optimal error estimates of the limiters are sharp by artificial examples in 2.1 and 2.2, in numerical tests we have not observed any degeneracy of orders of accuracy. Additionally, we have not observed any degeneracy of orders of accuracy when using the type-1 limiter with the $(k+1)$-point Gauss-Radau quadrature.

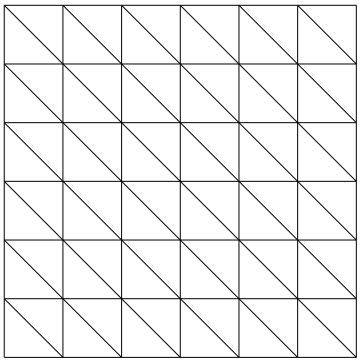


Figure 2: A typical triangular mesh used in the tests. Triangles are obtained by splitting rectangles diagonally.

Example 6.1. ***(Comparison of the conservation property for different limiters)***

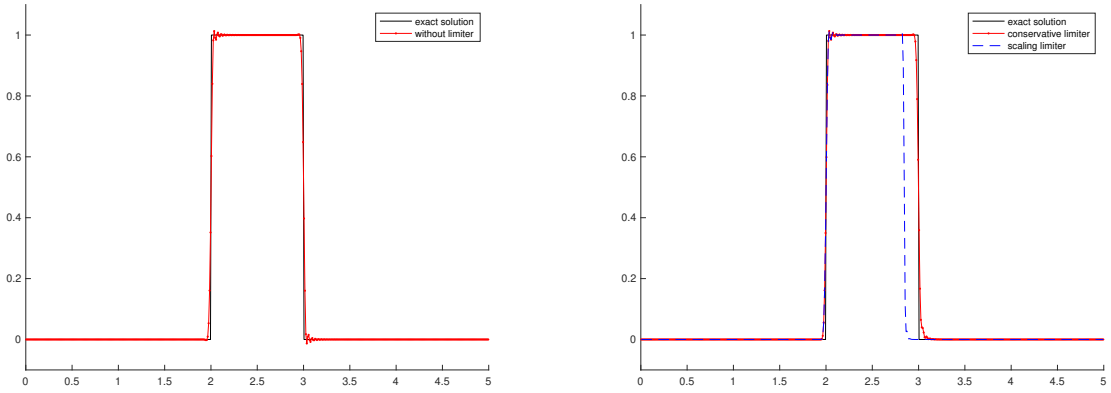
We solve the simple hyperbolic equation $u_t + u_x = 0$ with implicit time discretization by a variety of positivity-preserving schemes, and compare the results of different positivity-preserving limiters. We first compute the solutions using the scaling limiter [35] that preserves cell averages. Then, we replace the scaling limiter in these algorithms by our conservative limiter that preserves the sum of cell average and outflow fluxes. Since the only difference is in the use of limiters, it would be convincing that our notion of conservation is more appropriate if the results of the conservative limiters are better than those of the scaling limiter.

The initial and boundary condition are given below

$$u(x,0) = \begin{cases} 1, & 0 < x \le 1 \\ 0, & otherwise \end{cases}, \quad x \in \Omega; \quad u(0,t) = 0, \quad t \in [0,T].$$

We first compute the equation on the domain $\Omega = [0,5]$, with backward Euler time discretization, CFL number $\frac{\Delta t}{\Delta x} = 0.01$ and spatial partition $N = 500$, to the terminal time $T = 2$, based on the positivity-preserving P^2-DG scheme proposed in [30] for one dimensional linear equations. We plot the cell averages

of the numerical solutions at the terminal time for the cases with no limiter, with the scaling limiter and with the conservative limiter, and compare them with the exact solution. The results are shown in Figure 3, from which we can clearly observe a wrong shock location with the use of the scaling limiter. The total mass of the exact solution at the terminal time is $\int_\Omega u(x,T)dx = 1$. *In the numerical solutions, the total mass has changed* 3.10×10^{-12}, 2.94×10^{-12} *and* 1.54×10^{-1} *for the cases with no limiter, with the conservative limiter and with the scaling limiter, respectively.*

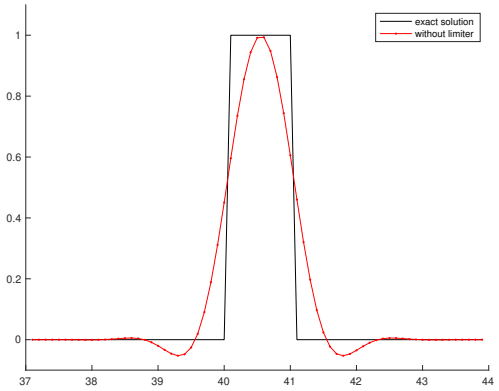


(a) No limiter

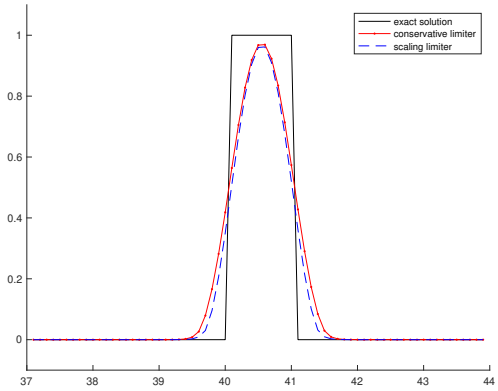
(b) With limiter

Figure 3: Comparison of results using different limiters in the P^2-DG scheme proposed in [30] for the equation $u_t + u_x = 0$ at $T = 2$ on the domain $\Omega = [0, 5]$ with backward Euler time discretization. The CFL number is set as $\frac{\Delta t}{\Delta x} = 0.01$ and the spatial partition is uniform with $N = 500$.

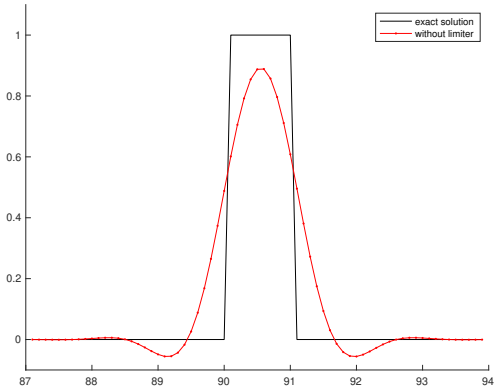
We then compute the equation on the space-time box $\Omega \times [0, T]$ *by the space-time DG discretization, based on the positivity-preserving* R^1*-DG scheme proposed in [21] for two dimensional linear equations. We take two space-time boxes* $\Omega_1 = [0, 50], T_1 = 40$ *and* $\Omega_2 = [0, 100], T_2 = 90$, *on the uniform meshes* $N_x^1 \times N_t^1 = 500 \times 400$ *and* $N_x^2 \times N_t^2 = 1000 \times 900$, *respectively. We plot the cell averages of the numerical solutions at the terminal times for the cases with no limiter, with the conservative limiter and with the scaling limiter, and compare them with the exact solution. The results are shown in Figure 4, from which we can clearly observe the loss of mass with the use of the scaling limiter. In the numerical solutions, the total mass in the domain has changed* 4.97×10^{-14}, 4.77×10^{-14} *and* 8.59×10^{-2} *at* T_1 *for the cases with no limiter, with the conservative limiter and with the scaling limiter, respectively, and* 1.01×10^{-13}, 1.11×10^{-13} *and* 1.47×10^{-1} *at* T_2 *for the cases with no limiter, with the conservative limiter and with the scaling limiter, respectively.*



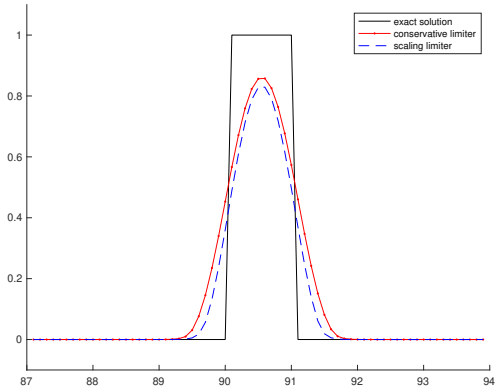
(a) $T_1 = 40$, no limiter



(b) $T_1 = 40$, with limiter



(c) $T_2 = 90$, no limiter



(d) $T_2 = 90$, with limiter

Figure 4: Comparison of results using different limiters in the R^1-DG scheme proposed in [21] for the equation $u_t + u_x = 0$ with space-time DG discretization. Solutions of (a), (b) are computed on $\Omega_1 \times [0, T_1] = [0, 50] \times [0, 40]$ with uniform partition $N_x^1 \times N_t^1 = 500 \times 400$. Solutions of (c), (d) are computed on $\Omega_2 \times [0, T_2] = [0, 100] \times [0, 90]$ with uniform partition $N_x^2 \times N_t^2 = 1000 \times 900$.

Finally, we compute the equation by the space-time DG based on the Q^2 and Q^3 schemes (3.2). *Note that these two schemes are not positivity-preserving for cell averages in general, namely, there is no theoretical guarantee that the cell averages always remain nonnegative with the use of the scaling limiter. However, by trial and error, we find a setting that keeps the cell averages nonnegative during simulation, with the use of positivity-preserving scaling limiter. We take the space-time box $\Omega = [0, 30], T = 25$ on the uniform mesh $N_x \times N_t = 300 \times 250$. We plot the cell averages of the numerical solutions at the terminal time for the cases with no limiter, with the conservative limiter and with the scaling limiter, and compare them with the exact solution. The results are shown in Figure 5, from which we can clearly observe the loss of mass with the use of the scaling limiter. The total mass in the domain have changed 8.48×10^{-14}, 8.53×10^{-14} and 1.88×10^{-1} in the Q^2-DG scheme for the cases with no limiter, with the conservative limiter and with the scaling limiter, respectively, and 1.87×10^{-13}, 1.88×10^{-13} and 7.78×10^{-2} in the Q^3-DG scheme for the cases with no limiter, with the conservative limiter and with the scaling limiter, respectively.*

Example 6.2. ***A linear stationary hyperbolic equation in one dimension***

We solve the equation (2.1) *with $a(x) = 1$, $\lambda = 6000$ and $s(x) = \lambda\left(\frac{1}{9}\cos^4(x) + \epsilon\right) - \frac{4}{9}\cos^3(x)\sin(x)$ on the domain $\Omega = [0, \pi]$, where $\epsilon = 10^{-14}$ is taken such that the source term is nonnegative. The boundary condition of the problem is $u(0) = \frac{1}{9} + \epsilon$ and the exact solution is $u(x) = \frac{1}{9}\cos^4(x) + \epsilon$.*

We compute the equation using the P^k-DG scheme (2.2) *with $k = 1, 2, 3, 4$. The errors, orders of convergence and data about positivity are given in Table 1, where we can observe that the negative values of the original scheme are eliminated by the limiter and the order of accuracy remains optimal.*

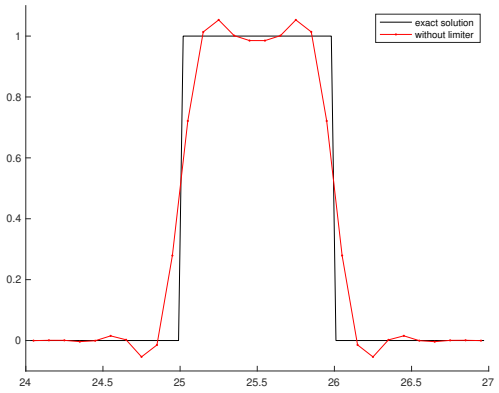
Example 6.3. ***A nonlinear stationary hyperbolic equation in one dimension***

We solve the equation (5.1) *with $f(u) = u^3 + 0.01u$, $\lambda = 5$ and $s(x) = -8\sin(x)\cos^7(x)\left(3(\cos^8(x) + \epsilon)^2 + 0.01\right) + \lambda\left(\cos^8(x) + \epsilon\right)$ on the domain $\Omega = [0, \pi]$, where $\epsilon = 10^{-14}$ is taken such that the source term is nonnegative. The boundary condition of the problem is $u(0) = 1 + \epsilon$ and the exact solution is $u(x) = \cos^8(x) + \epsilon$.*

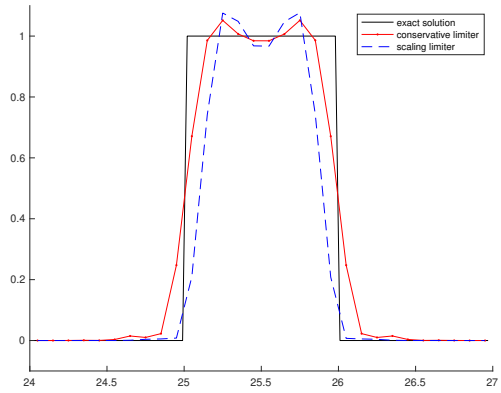
We compute the equation using the P^k-DG scheme (5.2) *with $k = 1, 2, 3, 4$. The errors, orders of convergence and data about positivity are given in Table 2. Same to the linear case, we can observe that the negative values of the original scheme are eliminated by the limiter and the order of accuracy remains optimal.*

Example 6.4. ***A nonlinear time-dependent hyperbolic equation in one dimension with backward Euler time discretization***

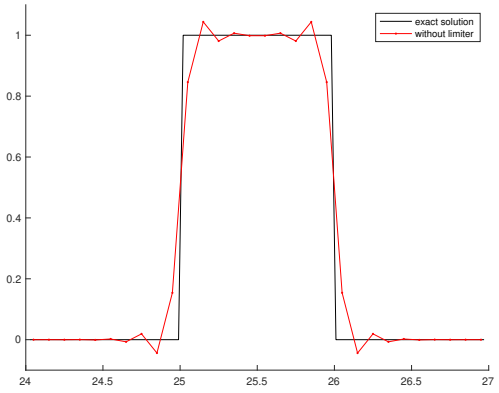
We solve the equation (1.1) *with backward Euler time discretization, and take $f(u) = \frac{u^3}{3}$, $s(x) = 0$. The*



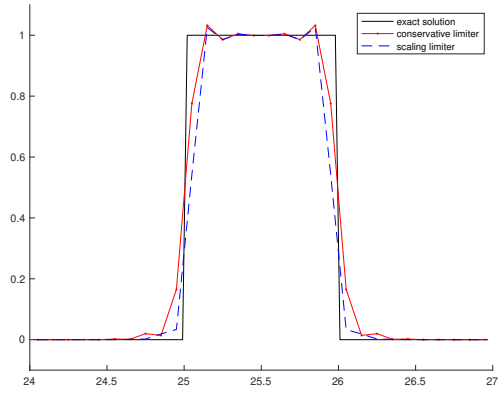
(a) $T = 25$, Q^2-DG, no limiter



(b) $T = 25$, Q^2-DG, with limiter



(c) $T = 25$, Q^3-DG, no limiter



(d) $T = 25$, Q^3-DG, with limiter

Figure 5: Comparison of results using different limiters in the scheme (3.2) for the equation $u_t + u_x = 0$ with space-time DG discretization. Solutions of are computed on $\Omega \times [0, T] = [0, 30] \times [0, 25]$ with uniform partition $N_x \times N_t = 300 \times 250$.

		no Limiter					with Limiter				
k	N	L^1 error	order	L^∞ error	order	$\min u_h$	L^1 error	order	L^∞ error	order	Limited cells (%)
1	20	4.62E-04	-	9.00E-04	-	-4.67E-05	4.63E-04	-	9.00E-04	-	20.00
	40	1.16E-04	1.99	2.28E-04	1.98	-3.34E-06	1.16E-04	2.00	2.28E-04	1.98	10.00
	80	2.90E-05	2.00	5.83E-05	1.97	-2.19E-07	2.90E-05	2.00	5.83E-05	1.97	5.00
	160	7.26E-06	2.00	1.50E-05	1.96	-1.43E-08	7.26E-06	2.00	1.50E-05	1.96	2.50
	320	1.82E-06	2.00	3.90E-06	1.94	-9.54E-10	1.82E-06	2.00	3.90E-06	1.94	1.25
2	20	2.04E-05	-	3.88E-05	-	-2.30E-06	2.05E-05	-	3.88E-05	-	10.00
	40	2.54E-06	3.01	4.84E-06	3.00	-1.49E-07	2.54E-06	3.01	4.84E-06	3.00	5.00
	80	3.19E-07	2.99	5.98E-07	3.02	-9.58E-09	3.19E-07	2.99	5.98E-07	3.02	2.50
	160	4.01E-08	2.99	7.33E-08	3.03	-6.28E-10	4.01E-08	2.99	7.33E-08	3.03	1.25
	320	5.09E-09	2.98	8.86E-09	3.05	-4.25E-11	5.09E-09	2.98	8.86E-09	3.05	0.63
3	20	7.72E-07	-	1.57E-06	-	-9.43E-07	9.58E-07	-	4.24E-06	-	10.00
	40	4.79E-08	4.01	1.03E-07	3.93	-6.21E-08	5.38E-08	4.16	2.76E-07	3.94	5.00
	80	3.01E-09	3.99	6.74E-09	3.93	-4.05E-09	3.19E-09	4.07	1.73E-08	4.00	2.50
	160	1.89E-10	4.00	4.48E-10	3.91	-2.69E-10	1.94E-10	4.04	1.06E-09	4.02	1.25
	320	1.19E-11	3.98	3.06E-11	3.87	-1.83E-11	1.21E-11	4.00	6.46E-11	4.04	0.63
4	20	2.44E-08	-	4.80E-08	-	-1.12E-08	2.71E-08	-	4.80E-08	-	10.00
	40	7.60E-10	5.01	1.47E-09	5.03	-1.83E-10	7.81E-10	5.12	1.47E-09	5.03	5.00
	80	2.39E-11	4.99	4.45E-11	5.04	-3.02E-12	2.41E-11	5.02	4.45E-11	5.04	2.50
	160	7.57E-13	4.98	1.32E-12	5.07	-4.17E-14	7.58E-13	4.99	1.32E-12	5.07	1.25
	320	2.43E-14	4.96	3.97E-14	5.06	9.10E-15	2.43E-14	4.96	3.97E-14	5.06	0.00

Table 1: Results for the equation $u_x + 6000u = 6000(\frac{1}{9}\cos^4(x) + 10^{-14}) - \frac{4}{9}\cos^3(x)\sin(x)$ in Example 6.2 using the P^k-DG scheme (2.2) on the domain $\Omega = [0, \pi]$

		no Limiter					with Limiter				
k	N	L^1 error	order	L^∞ error	order	$\min u_h$	L^1 error	order	L^∞ error	order	Limited cells (%)
1	20	7.45E-03	-	3.02E-02	-	-1.49E-04	7.45E-03	-	3.02E-02	-	30.00
	40	1.91E-03	1.96	7.97E-03	1.92	-8.48E-07	1.91E-03	1.96	7.97E-03	1.92	17.50
	80	4.91E-04	1.96	2.03E-03	1.97	-6.90E-09	4.91E-04	1.96	2.03E-03	1.97	8.75
	160	1.26E-04	1.96	5.11E-04	1.99	-5.30E-11	1.26E-04	1.96	5.11E-04	1.99	3.13
	320	3.22E-05	1.97	1.28E-04	2.00	-3.36E-13	3.22E-05	1.97	1.28E-04	2.00	0.63
2	20	4.57E-04	-	2.12E-03	-	-1.27E-06	4.57E-04	-	2.12E-03	-	20.00
	40	5.72E-05	3.00	2.62E-04	3.01	-1.18E-08	5.72E-05	3.00	2.62E-04	3.01	10.00
	80	7.22E-06	2.98	3.20E-05	3.03	-7.14E-11	7.22E-06	2.98	3.20E-05	3.03	5.00
	160	9.17E-07	2.98	3.96E-06	3.01	-2.48E-13	9.17E-07	2.98	3.96E-06	3.01	1.88
	320	1.16E-07	2.98	4.92E-07	3.01	9.44E-15	1.16E-07	2.98	4.92E-07	3.01	0.00
3	20	2.30E-05	-	1.19E-04	-	-9.63E-07	2.30E-05	-	1.19E-04	-	20.00
	40	1.44E-06	3.99	7.83E-06	3.93	-5.10E-09	1.44E-06	3.99	7.83E-06	3.93	10.00
	80	9.24E-08	3.97	4.95E-07	3.98	-2.42E-11	9.24E-08	3.97	4.95E-07	3.98	5.00
	160	5.86E-09	3.98	3.11E-08	3.99	-1.07E-13	5.86E-09	3.98	3.11E-08	3.99	0.63
	320	3.71E-10	3.98	1.94E-09	4.00	9.43E-15	3.71E-10	3.98	1.94E-09	4.00	0.00
4	20	1.05E-06	-	6.46E-06	-	-6.21E-09	1.05E-06	-	6.46E-06	-	10.00
	40	3.31E-08	4.98	1.88E-07	5.10	-3.05E-11	3.31E-08	4.98	1.88E-07	5.10	5.00
	80	1.05E-09	4.97	5.80E-09	5.02	-8.17E-14	1.05E-09	4.97	5.80E-09	5.02	2.50
	160	3.35E-11	4.98	1.79E-10	5.02	9.65E-15	3.35E-11	4.98	1.79E-10	5.02	0.00
	320	1.06E-12	4.98	5.54E-12	5.01	1.00E-14	1.06E-12	4.98	5.54E-12	5.01	0.00

Table 2: Results for the equation $(u^3 + 0.01u)_x + 5u = -8\sin(x)\cos^7(x)\left(3(\cos^8(x) + 10^{-14})^2 + 0.01\right) + 5\left(\cos^8(x) + 10^{-14}\right)$ in Example 6.3 using the P^k-DG scheme (5.2) on the domain $\Omega = [0, \pi]$

initial and boundary condition of the equation are given below

$$u(x,0) = \begin{cases} 1, & x \le 1 \\ \\ 0, & otherwise \end{cases}, \quad x \in \Omega; \quad u(0,t) = 1, \quad t \in [0,T],$$

where $\Omega = [0,3]$ *and* $T = 2.5$.

We compute the equation using the P^k*-DG scheme* (5.2) *with* $k = 1,2,3,4$*, CFL number* $\frac{\Delta t}{\Delta x} = 0.5$ *and spatial partition* $N = 150$*. We zoom in the pre-shock zone and draw the cell averages of the numerical solutions in this area in Figure 6, with a comparison with the exact solution and the results without limiter. From the figures, we can observe that the negative cell averages of the original numerical scheme are eliminated by the limiter.*

Example 6.5. ***Linear stationary hyperbolic equations in two dimensions with smooth solutions***

We solve the equation (3.1) *with constant coefficients* $a(x,y) \equiv a = 0.7, b(x,y) \equiv b = 0.3$ *and* $s(x,y) \equiv 0$ *in the domain* $\Omega = [0,1] \times [0,1]$*. The inflow boundary conditions are* $u(x,0) = 0, 0 < x \le 1$ *and* $u(0,y) = \sin^6(\pi y), 0 \le y \le 1$*. The exact solution of the problem is*

$$u(x,y) = \begin{cases} 0, & y < \frac{b}{a}x \\ \\ \sin^6(\pi(y - \frac{b}{a}x))e^{-\frac{\lambda}{a}x} & y \ge \frac{b}{a}x. \end{cases}$$

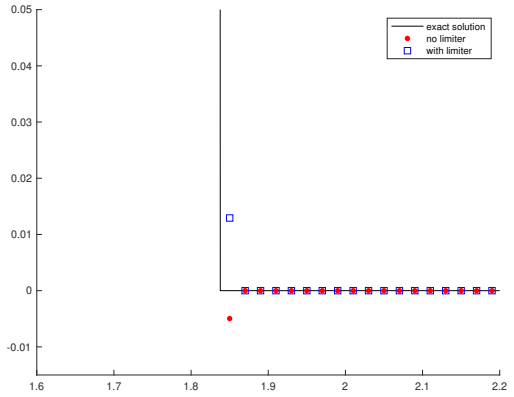
We take $\lambda = 1$*, which corresponds to the purely absorbing medium in RTE, and* $\lambda = 0$*, which corresponds to the transparent medium in RTE, in the tests.*

We compute the equations using the Q^k*-DG scheme* (3.2) *on rectangular meshes, and the* P^k*-DG scheme* (4.1) *on triangular meshes. The errors, orders of convergence and data about positivity are given in Table 3 - Table 6, from which we can observe that the positivity and optimal accuracy are both attained by the algorithms.*

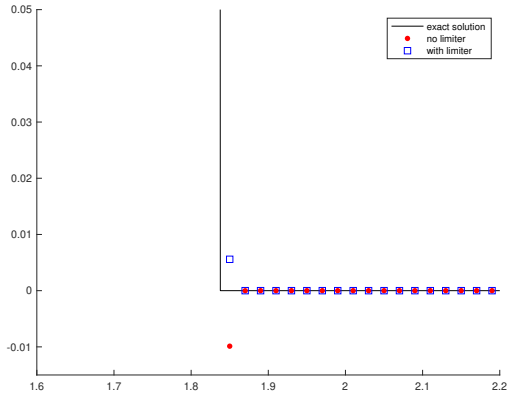
Example 6.6. ***Linear stationary hyperbolic equations in two dimensions with discontinuous solutions***

We solve the equation (3.1) *with constant coefficients* $a(x,y) \equiv a = 0.7, b(x,y) \equiv b = 0.3, s(x,y) \equiv 0$ *in the domain* $\Omega = [0,1] \times [0,1]$*. The inflow boundary conditions are* $u(x,0) = 0, 0 < x \le 1$ *and* $u(0,y) = 1, 0 \le y \le 1$*. The exact solution of the problem is*

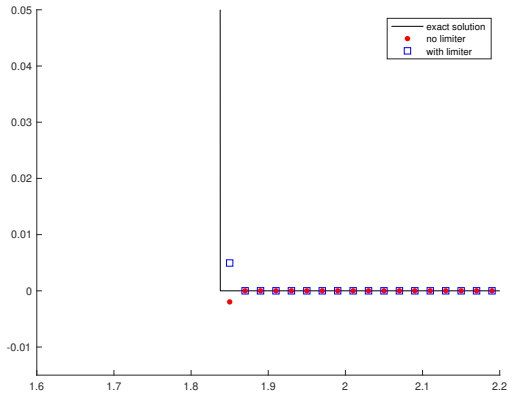
$$u(x,y) = \begin{cases} 0, & y < \frac{b}{a}x \\ \\ e^{-\frac{\lambda}{a}x} & y \ge \frac{b}{a}x. \end{cases}$$



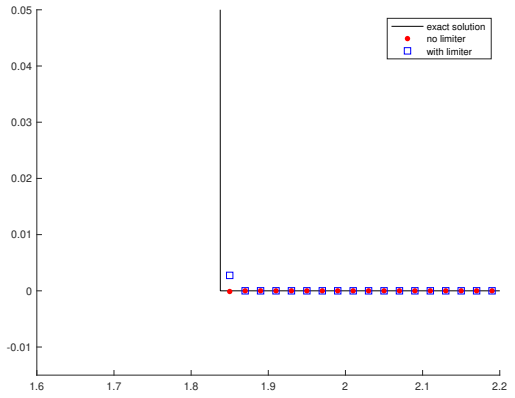
(a) P^1-DG



(b) P^2-DG



(c) P^3-DG



(d) P^4-DG

Figure 6: A zoomed-in view of the results for the equation $u_t + \left(\frac{u^3}{3}\right)_x = 0$ in Example 6.4 using the P^k-DG scheme (5.2) at $T = 2.5$ on the domain $\Omega = [0, 3]$ with backward Euler time discretization. The CFL number is set as $\frac{\Delta t}{\Delta x} = 0.5$ and the spatial partition is uniform with $N = 150$.

		no Limiter					with Limiter				
k	$N_x \times N_y$	L^1 error	order	L^∞ error	order	$\min u_h$	L^1 error	order	L^∞ error	order	Limited cells (%)
1	20×20	1.43E-03	-	2.66E-02	-	-3.07E-04	1.42E-03	-	2.66E-02	-	35.00
	40×40	3.38E-04	2.08	6.94E-03	1.94	-2.06E-05	3.38E-04	2.07	6.94E-03	1.94	29.25
	80×80	8.21E-05	2.04	1.76E-03	1.98	-1.03E-06	8.21E-05	2.04	1.76E-03	1.98	25.98
	160×160	2.03E-05	2.02	4.42E-04	1.99	-4.61E-08	2.03E-05	2.02	4.42E-04	1.99	24.05
	320×320	5.04E-06	2.01	1.11E-04	2.00	-1.99E-09	5.04E-06	2.01	1.11E-04	2.00	22.95
2	20×20	6.32E-05	-	1.26E-03	-	-1.59E-06	6.32E-05	-	1.26E-03	-	14.50
	40×40	7.79E-06	3.02	1.63E-04	2.96	-2.73E-08	7.79E-06	3.02	1.63E-04	2.96	13.13
	80×80	9.71E-07	3.01	2.07E-05	2.98	-4.36E-10	9.71E-07	3.01	2.07E-05	2.98	12.58
	160×160	1.21E-07	3.00	2.60E-06	2.99	-6.86E-12	1.21E-07	3.00	2.60E-06	2.99	12.23
	320×320	1.51E-08	3.00	3.26E-07	3.00	-1.15E-13	1.51E-08	3.00	3.26E-07	3.00	12.06
3	20×20	2.77E-06	-	6.61E-05	-	-3.64E-07	2.78E-06	-	6.61E-05	-	18.50
	40×40	1.72E-07	4.01	4.35E-06	3.92	-7.14E-09	1.72E-07	4.01	4.35E-06	3.92	16.56
	80×80	1.07E-08	4.00	2.76E-07	3.98	-1.21E-10	1.07E-08	4.00	2.76E-07	3.98	15.55
	160×160	6.71E-10	4.00	1.73E-08	3.99	-1.94E-12	6.71E-10	4.00	1.73E-08	3.99	15.20
	320×320	4.19E-11	4.00	1.08E-09	4.00	-3.06E-14	4.19E-11	4.00	1.08E-09	4.00	14.98
4	20×20	1.08E-07	-	2.56E-06	-	-2.03E-08	1.09E-07	-	2.56E-06	-	19.50
	40×40	3.37E-09	5.01	8.33E-08	4.94	-3.45E-10	3.38E-09	5.01	8.33E-08	4.94	17.94
	80×80	1.05E-10	5.00	2.62E-09	4.99	-5.55E-12	1.05E-10	5.00	2.62E-09	4.99	17.31
	160×160	3.29E-12	5.00	8.19E-11	5.00	-9.54E-14	3.29E-12	5.00	8.19E-11	5.00	16.75
	320×320	1.07E-13	4.94	2.57E-12	5.00	-2.15E-15	1.07E-13	4.94	2.57E-12	5.00	16.53

Table 3: Results for the equation $0.7u_x + 0.3u_y + u = 0$ in Example 6.5 using the Q^k-DG scheme (3.2) on rectangular meshes in domain $\Omega = [0, 1]^2$

		no Limiter					with Limiter				
k	$N_x \times N_y$	L^1 error	order	L^∞ error	order	$\min u_h$	L^1 error	order	L^∞ error	order	Limited cells (%)
1	20×20	1.92E-03	-	2.37E-02	-	-2.89E-03	1.89E-03	-	2.37E-02	-	22.13
	40×40	4.40E-04	2.13	6.85E-03	1.79	-1.32E-04	4.40E-04	2.10	6.85E-03	1.79	17.22
	80×80	1.05E-04	2.07	1.81E-03	1.92	-5.02E-06	1.05E-04	2.07	1.81E-03	1.92	14.23
	160×160	2.57E-05	2.03	4.67E-04	1.96	-1.88E-07	2.57E-05	2.03	4.67E-04	1.96	12.62
	320×320	6.38E-06	2.01	1.18E-04	1.98	-7.32E-09	6.38E-06	2.01	1.18E-04	1.98	11.77
2	20×20	9.43E-05	-	1.92E-03	-	-2.44E-05	9.47E-05	-	1.92E-03	-	10.38
	40×40	1.14E-05	3.05	2.62E-04	2.87	-4.59E-07	1.14E-05	3.05	2.62E-04	2.87	8.28
	80×80	1.42E-06	3.01	3.45E-05	2.92	-7.72E-09	1.42E-06	3.01	3.45E-05	2.92	7.74
	160×160	1.77E-07	3.00	4.39E-06	2.97	-1.23E-10	1.77E-07	3.00	4.39E-06	2.97	7.42
	320×320	2.21E-08	3.00	5.54E-07	2.99	-1.95E-12	2.21E-08	3.00	5.54E-07	2.99	7.34
3	20×20	5.87E-06	-	1.04E-04	-	-1.03E-05	6.77E-06	-	1.04E-04	-	14.75
	40×40	3.70E-07	3.99	7.30E-06	3.83	-2.32E-07	3.78E-07	4.16	7.30E-06	3.83	11.94
	80×80	2.31E-08	4.00	4.85E-07	3.91	-4.05E-09	2.32E-08	4.03	4.85E-07	3.91	10.50
	160×160	1.44E-09	4.00	3.13E-08	3.96	-6.52E-11	1.44E-09	4.00	3.13E-08	3.96	9.86
	320×320	9.02E-11	4.00	1.98E-09	3.98	-1.03E-12	9.02E-11	4.00	1.98E-09	3.98	9.43
4	20×20	3.16E-07	-	5.85E-06	-	-5.81E-07	3.32E-07	-	5.85E-06	-	9.75
	40×40	9.83E-09	5.01	2.05E-07	4.83	-1.08E-08	9.97E-09	5.06	2.05E-07	4.83	8.75
	80×80	3.03E-10	5.02	6.72E-09	4.93	-1.89E-10	3.04E-10	5.03	6.72E-09	4.93	8.59
	160×160	9.41E-12	5.01	2.21E-10	4.93	-3.08E-12	9.41E-12	5.01	2.21E-10	4.93	8.36
	320×320	3.06E-13	4.94	7.25E-12	4.93	-4.90E-14	3.06E-13	4.95	7.25E-12	4.93	8.31

Table 4: Results for the equation $0.7u_x + 0.3u_y + u = 0$ in Example 6.5 using the P^k-DG scheme (4.1) on triangular meshes in domain $\Omega = [0, 1]^2$

		no Limiter					with Limiter				
k	$N_x \times N_y$	L^1 error	order	L^∞ error	order	$\min u_h$	L^1 error	order	L^∞ error	order	Limited cells (%)
1	20×20	2.66E-03	-	2.81E-02	-	-1.12E-03	2.64E-03	-	2.81E-02	-	35.00
	40×40	6.10E-04	2.12	7.23E-03	1.96	-8.44E-05	6.10E-04	2.11	7.23E-03	1.96	29.25
	80×80	1.46E-04	2.06	1.82E-03	1.99	-4.12E-06	1.46E-04	2.06	1.82E-03	1.99	25.97
	160×160	3.59E-05	2.03	4.56E-04	2.00	-1.87E-07	3.59E-05	2.03	4.56E-04	2.00	24.05
	320×320	8.89E-06	2.01	1.14E-04	2.00	-8.21E-09	8.89E-06	2.01	1.14E-04	2.00	22.95
2	20×20	1.14E-04	-	1.32E-03	-	-1.65E-06	1.14E-04	-	1.32E-03	-	14.50
	40×40	1.40E-05	3.03	1.67E-04	2.98	-2.78E-08	1.40E-05	3.03	1.67E-04	2.98	13.19
	80×80	1.74E-06	3.01	2.09E-05	3.00	-4.39E-10	1.74E-06	3.01	2.09E-05	3.00	12.56
	160×160	2.17E-07	3.00	2.61E-06	3.00	-1.32E-11	2.17E-07	3.00	2.61E-06	3.00	12.21
	320×320	2.71E-08	3.00	3.27E-07	3.00	-4.61E-13	2.71E-08	3.00	3.27E-07	3.00	12.07
3	20×20	4.96E-06	-	7.10E-05	-	-4.14E-07	4.97E-06	-	7.10E-05	-	19.25
	40×40	3.08E-07	4.01	4.48E-06	3.98	-7.61E-09	3.08E-07	4.01	4.48E-06	3.98	16.69
	80×80	1.92E-08	4.00	2.81E-07	4.00	-1.24E-10	1.92E-08	4.00	2.81E-07	4.00	15.53
	160×160	1.20E-09	4.00	1.76E-08	4.00	-3.10E-12	1.20E-09	4.00	1.76E-08	4.00	15.21
	320×320	7.50E-11	4.00	1.10E-09	4.00	-8.31E-14	7.50E-11	4.00	1.10E-09	4.00	15.01
4	20×20	1.94E-07	-	2.68E-06	-	-2.70E-08	1.95E-07	-	2.68E-06	-	19.25
	40×40	6.04E-09	5.01	8.42E-08	4.99	-5.42E-10	6.05E-09	5.01	8.42E-08	4.99	17.88
	80×80	1.88E-10	5.00	2.63E-09	5.00	-1.18E-11	1.88E-10	5.00	2.63E-09	5.00	17.30
	160×160	5.88E-12	5.00	8.24E-11	5.00	-2.73E-13	5.88E-12	5.00	8.24E-11	5.00	16.74
	320×320	1.92E-13	4.93	2.78E-12	4.89	-6.56E-15	1.92E-13	4.93	2.80E-12	4.88	16.48

Table 5: Results for the equation $0.7u_x + 0.3u_y = 0$ in Example 6.5 using the Q^k-DG scheme (3.2) on rectangular meshes in domain $\Omega = [0, 1]^2$

		no Limiter					with Limiter				
k	$N_x \times N_y$	L^1 error	order	L^∞ error	order	$\min u_h$	L^1 error	order	L^∞ error	order	Limited cells (%)
1	20×20	3.67E-03	-	2.70E-02	-	-5.59E-03	3.58E-03	-	2.70E-02	-	22.38
	40×40	8.02E-04	2.19	7.41E-03	1.87	-4.09E-04	8.02E-04	2.16	7.41E-03	1.87	17.22
	80×80	1.86E-04	2.11	1.91E-03	1.96	-1.85E-05	1.87E-04	2.10	1.91E-03	1.96	14.21
	160×160	4.52E-05	2.04	4.80E-04	1.99	-7.38E-07	4.52E-05	2.04	4.80E-04	1.99	12.61
	320×320	1.12E-05	2.02	1.20E-04	2.00	-2.94E-08	1.12E-05	2.02	1.20E-04	2.00	11.77
2	20×20	1.68E-04	-	2.05E-03	-	-2.81E-05	1.69E-04	-	2.05E-03	-	11.63
	40×40	2.02E-05	3.05	2.64E-04	2.96	-4.80E-07	2.02E-05	3.06	2.64E-04	2.96	9.25
	80×80	2.50E-06	3.02	3.32E-05	2.99	-7.89E-09	2.50E-06	3.02	3.32E-05	2.99	8.54
	160×160	3.12E-07	3.00	4.16E-06	3.00	-1.25E-10	3.12E-07	3.00	4.16E-06	3.00	8.27
	320×320	3.89E-08	3.00	5.21E-07	3.00	-1.96E-12	3.89E-08	3.00	5.21E-07	3.00	8.18
3	20×20	1.02E-05	-	1.14E-04	-	-1.22E-05	1.19E-05	-	1.14E-04	-	15.38
	40×40	6.43E-07	3.99	7.90E-06	3.85	-2.52E-07	6.58E-07	4.17	7.90E-06	3.85	12.50
	80×80	4.02E-08	4.00	5.10E-07	3.95	-4.20E-09	4.03E-08	4.03	5.10E-07	3.95	11.00
	160×160	2.51E-09	4.00	3.21E-08	3.99	-6.67E-11	2.51E-09	4.00	3.21E-08	3.99	10.34
	320×320	1.57E-10	4.00	2.01E-09	4.00	-1.05E-12	1.57E-10	4.00	2.01E-09	4.00	9.87
4	20×20	5.55E-07	-	6.22E-06	-	-6.07E-07	5.85E-07	-	6.22E-06	-	10.13
	40×40	1.71E-08	5.02	2.16E-07	4.85	-1.19E-08	1.73E-08	5.08	2.16E-07	4.85	9.34
	80×80	5.26E-10	5.02	6.96E-09	4.96	-1.99E-10	5.28E-10	5.04	6.96E-09	4.96	8.84
	160×160	1.63E-11	5.01	2.20E-10	4.98	-3.16E-12	1.63E-11	5.01	2.20E-10	4.98	8.48
	320×320	5.36E-13	4.93	7.06E-12	4.96	-4.96E-14	5.36E-13	4.93	7.06E-12	4.96	8.32

Table 6: Results for the equation $0.7u_x + 0.3u_y = 0$ in Example 6.5 using the P^k-DG scheme (4.1) on triangular meshes in domain $\Omega = [0, 1]^2$

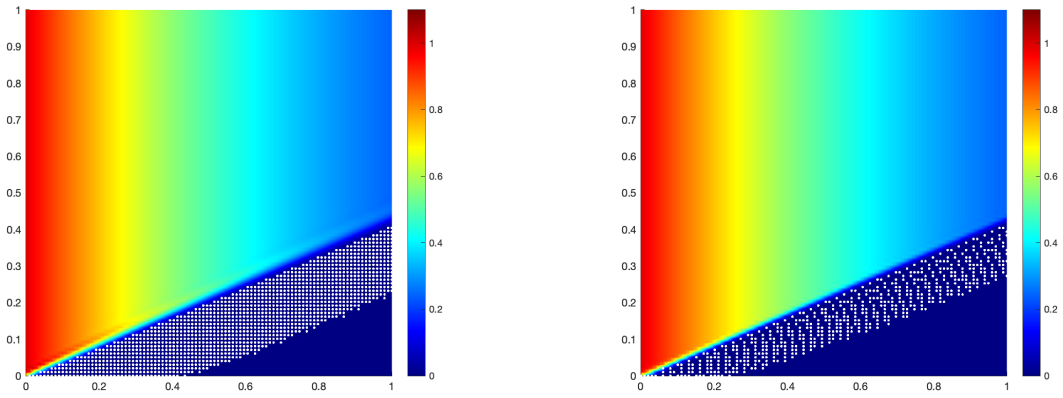
We test the cases of $\lambda = 1$ and $\lambda = 0$, which correspond to the purely absorbing medium and transparent medium in RTE, respectively.

The solutions are computed by the Q^k-DG scheme (3.2) *on rectangular meshes, and by the P^k-DG scheme* (4.1) *on triangular meshes, with the spatial partition $N_x \times N_y = 100 \times 100$. We draw the contours of the solutions on rectangular meshes in Figures 7 and 9 for the cases $\lambda = 1$ and $\lambda = 0$, respectively. The contours of the solutions on triangular meshes are given in Figures 11 and 13 for the cases $\lambda = 1$ and $\lambda = 0$, respectively. Moreover, we slice the solutions along $y = 0.25$ and plot the averages of the solution along the line in Figures 8, 10, 12 and 14. From the figures, we can observe that the negative averages of the solution in the original scheme are eliminated by the positivity-preserving technique.*

7 Concluding remarks

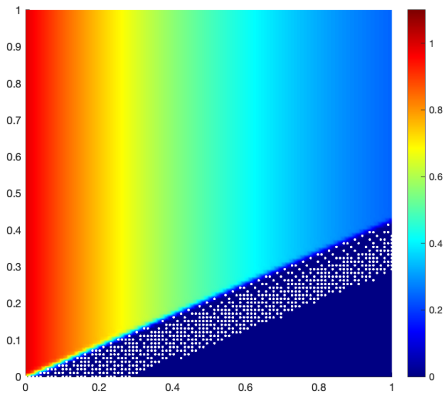
In this paper, we have constructed high order conservative positivity-preserving discontinuous Galerkin methods for various stationary hyperbolic equations in one and two space dimensions, based on a novel definition of conservation for stationary equations. Two types of conservative positivity-preserving limiters are introduced, where the type-1 limiter relies on particular Gauss-Radau quadratures for the schemes while the type-2 limiter does not. The errors introduced by the limiters are of optimal order on downstream edges, thus the limiter does not pollute from the original high order accuracy on downstream cells. Moreover, for time-dependent hyperbolic problems with implicit time discretization, the errors introduced by limiters are always optimal.

The positivity-preserving technique proposed in this paper is easy to implement, simple to prove for the positivity, and applicable for general types of stationary hyperbolic equations, compared with the previous works. Moreover, it is possible to apply this positivity-preserving approach to certain time-dependent advection systems using implicit discretization for target variables and explicit discretization for velocities in the advection terms. In the future, we will extend the methods to the bound-preserving DG for the multi-phase flow system and the positivity-preserving DG for the multi-components gas flow system in porous media in two space dimensions with implicit time discretization, based on the this methodology.

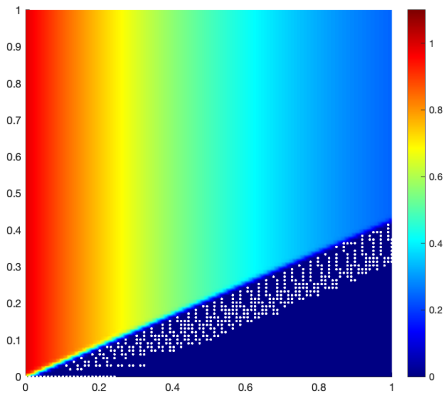


(a) $k = 1$

(b) $k = 2$

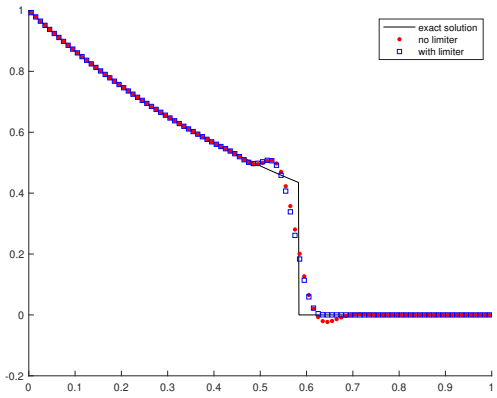


(c) $k = 3$

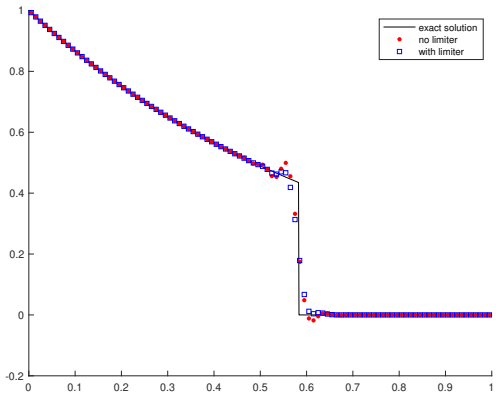


(d) $k = 4$

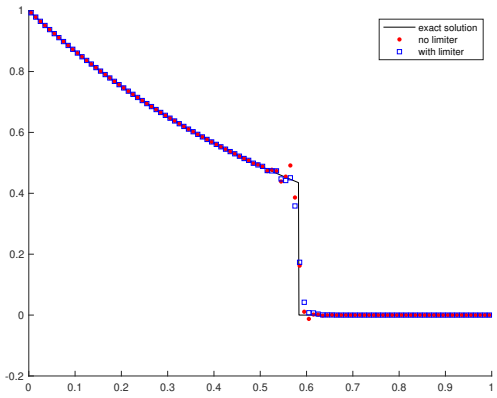
Figure 7: Solutions of $0.7u_x + 0.3u_y + u = 0$ in Example 6.6 using the Q^k-DG scheme (3.2) on rectangular meshes in domain $\Omega = [0, 1]^2$ with uniform partition $N_x \times N_y = 100 \times 100$.



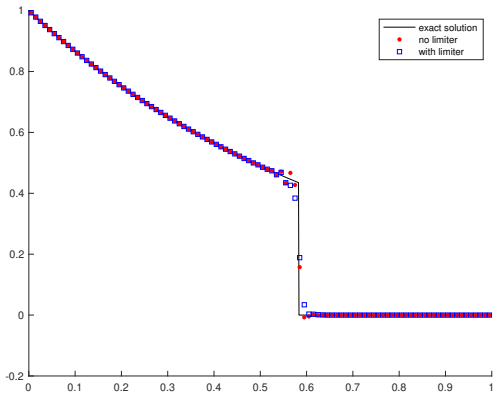
(a) $k = 1$



(b) $k = 2$

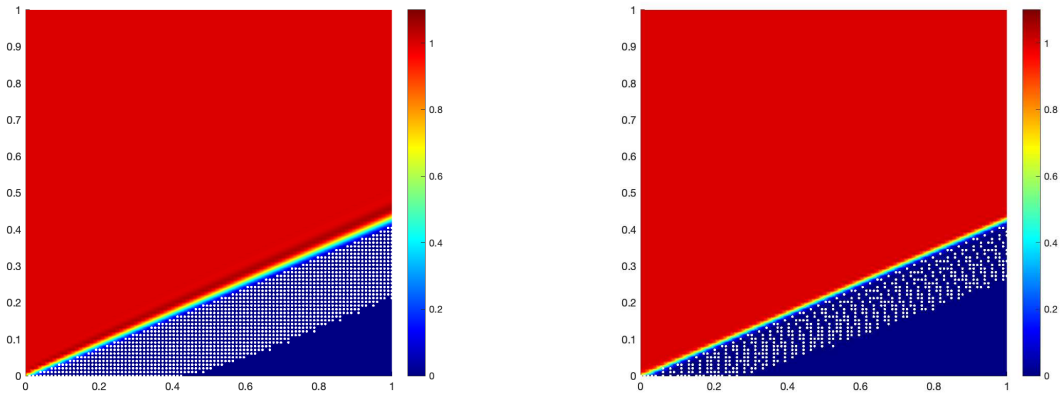


(c) $k = 3$



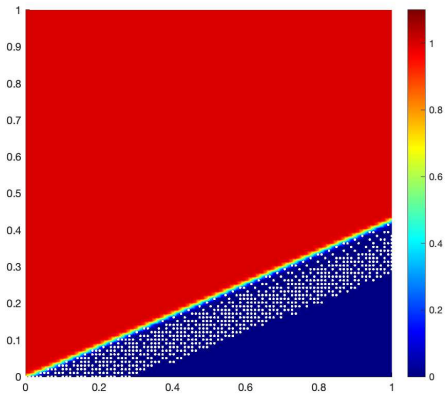
(d) $k = 4$

Figure 8: Solutions of $0.7u_x + 0.3u_y + u = 0$ in Example 6.6 using the Q^k-DG scheme (3.2) on rectangular meshes in domain $\Omega = [0, 1]^2$ with uniform partition $N_x \times N_y = 100 \times 100$, cut along $y = 0.25$

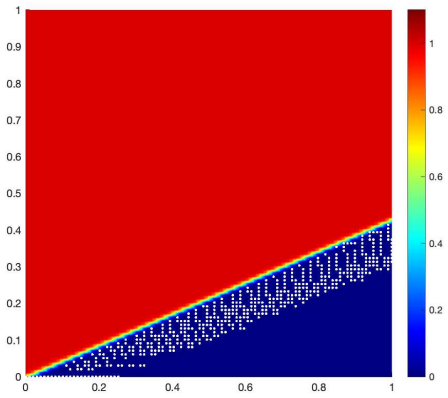


(a) $k = 1$

(b) $k = 2$

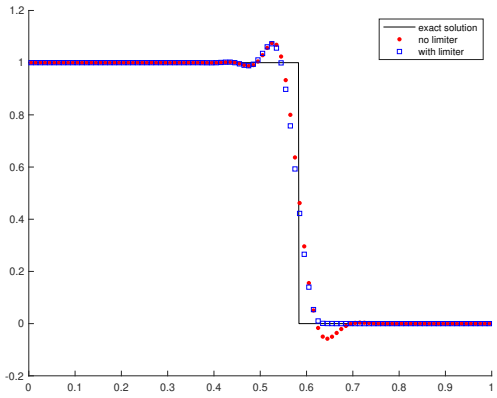


(c) $k = 3$

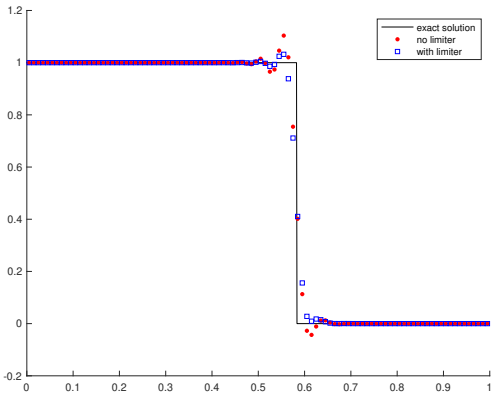


(d) $k = 4$

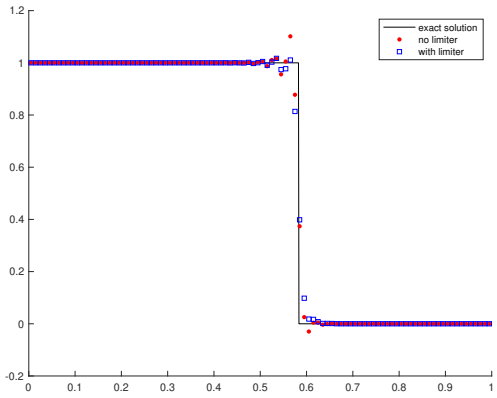
Figure 9: Solutions of $0.7u_x + 0.3u_y = 0$ in Example 6.6 using the Q^k-DG scheme (3.2) on rectangular meshes in domain $\Omega = [0, 1]^2$ with uniform partition $N_x \times N_y = 100 \times 100$.



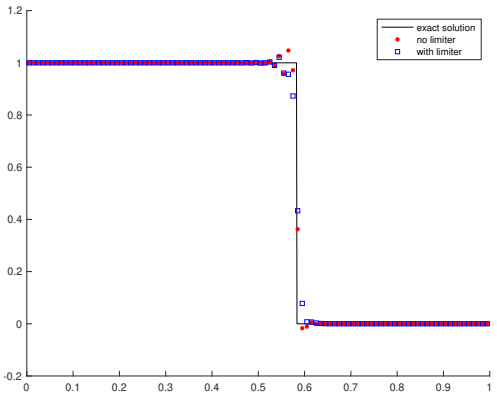
(a) $k = 1$



(b) $k = 2$

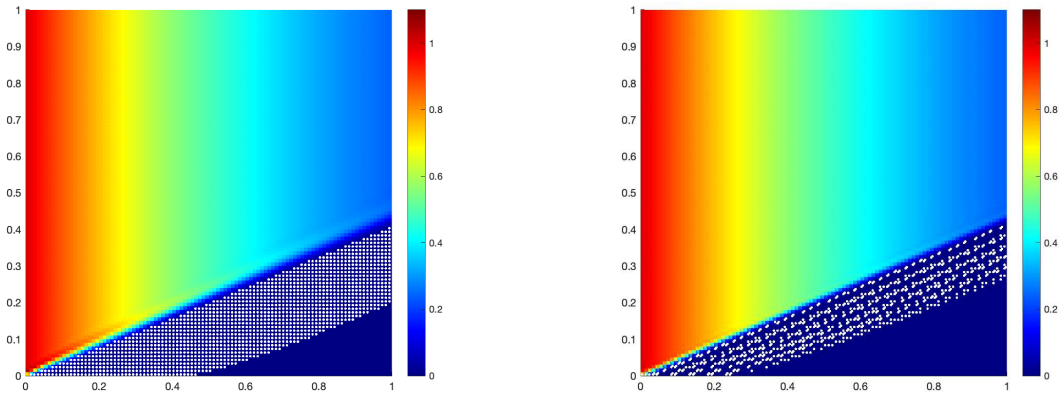


(c) $k = 3$



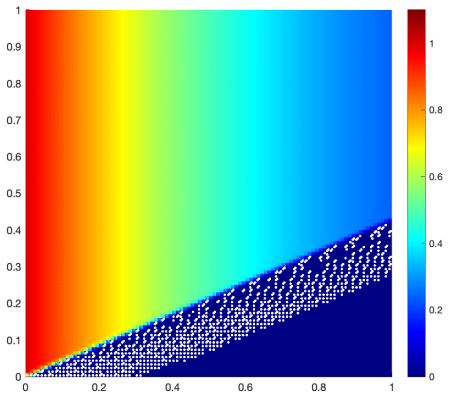
(d) $k = 4$

Figure 10: Solutions of $0.7u_x + 0.3u_y = 0$ in Example 6.6 using the Q^k-DG scheme (3.2) on rectangular meshes in domain $\Omega = [0, 1]^2$ with uniform partition $N_x \times N_y = 100 \times 100$, cut along $y = 0.25$

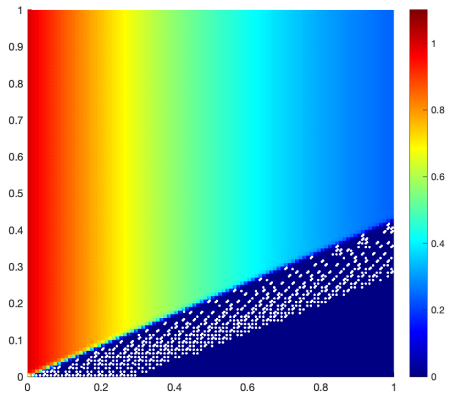


(a) $k = 1$

(b) $k = 2$

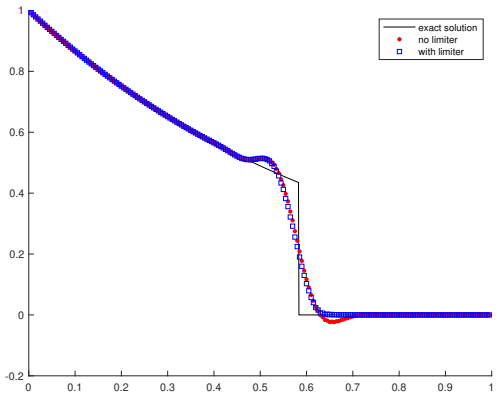


(c) $k = 3$

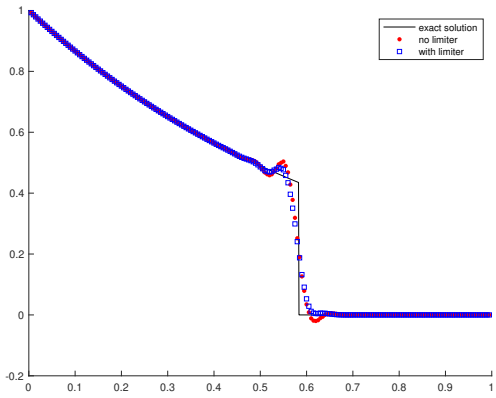


(d) $k = 4$

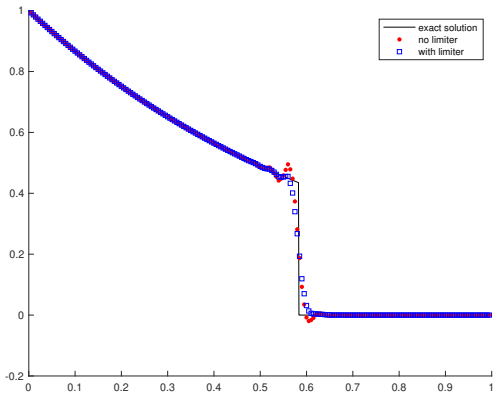
Figure 11: Solutions of $0.7u_x + 0.3u_y + u = 0$ in Example 6.6 using the P^k-DG scheme (4.1) on triangular meshes in domain $\Omega = [0, 1]^2$ with uniform partition $N_x \times N_y = 100 \times 100$.



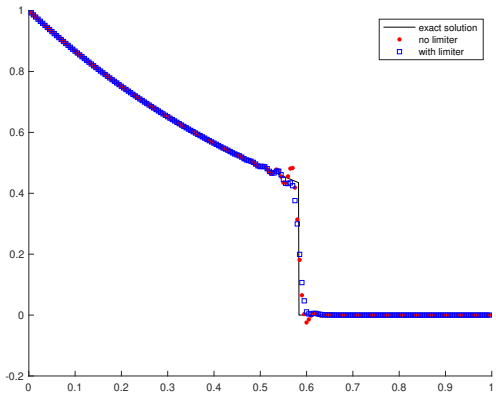
(a) $k = 1$



(b) $k = 2$

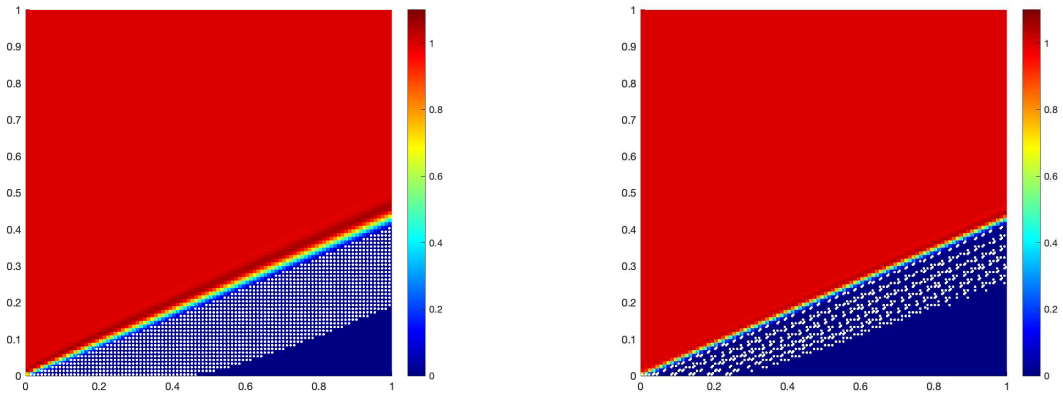


(c) $k = 3$



(d) $k = 4$

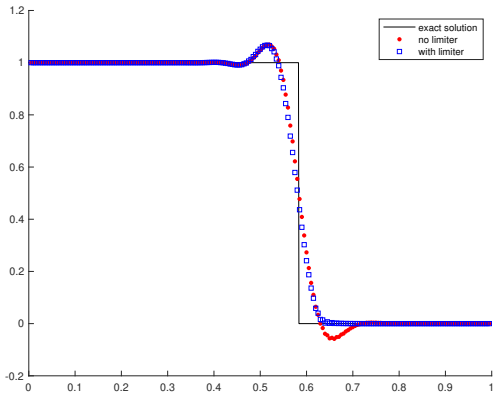
Figure 12: Solutions of $0.7u_x + 0.3u_y + u = 0$ in Example 6.6 using the P^k-DG scheme (4.1) on triangular meshes in domain $\Omega = [0, 1]^2$ with uniform partition $N_x \times N_y = 100 \times 100$, cut along $y = 0.25$



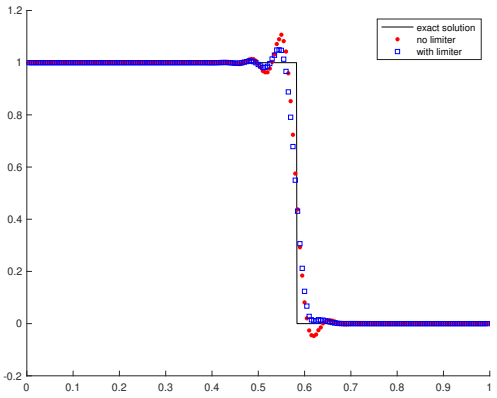
(a) $k = 1$ (b) $k = 2$

(c) $k = 3$ (d) $k = 4$

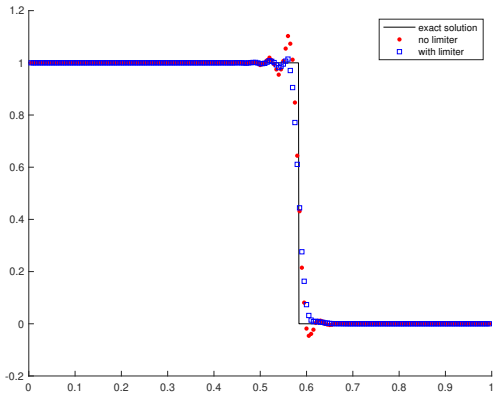
Figure 13: Solutions of $0.7u_x + 0.3u_y = 0$ in Example 6.6 using the P^k-DG scheme (4.1) on triangular meshes in domain $\Omega = [0, 1]^2$ with uniform partition $N_x \times N_y = 100 \times 100$.



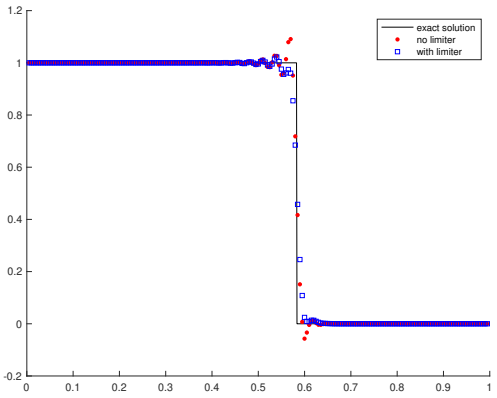
(a) $k = 1$



(b) $k = 2$



(c) $k = 3$



(d) $k = 4$

Figure 14: Solutions of $0.7u_x + 0.3u_y = 0$ in Example 6.6 using the P^k-DG scheme (4.1) on triangular meshes in domain $\Omega = [0, 1]^2$ with uniform partition $N_x \times N_y = 100 \times 100$, cut along $y = 0.25$

References

[1] R. Anderson, V. Dobrev, T. Kolev, D. Kuzmin, M. Quezada de Luna, R. Rieben and V. Tomov, High-order local maximum principle preserving (MPP) discontinuous Galerkin finite element method for the transport equation, Journal of Computational Physics, 334, 2017, 102-124.

[2] T. Chen and C.-W. Shu, Entropy stable high order discontinuous Galerkin methods with suitable quadrature rules for hyperbolic conservation laws, Journal of Computational Physics, 345, 2017, 427-461.

[3] Z. Chen, H. Huang and J. Yan, Third order maximum-principle-satisfying direct discontinuous Galerkin methods for time dependent convection diffusion equations on unstructured triangular meshes, Journal of Computational Physics, 308, 2016, 198-217.

[4] N. Chuenjarern, Z. Xu and Y. Yang, High-order bound-preserving discontinuous Galerkin methods for compressible miscible displacements in porous media on triangular meshes, Journal of Computational Physics, 378, 2019, 110-128.

[5] B. Cockburn, S. Hou and C.-W. Shu, The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: the multidimensional case, Mathematics of Computation, 54, 1990, 545-581.

[6] B. Cockburn, S.-Y. Lin and C.-W. Shu, TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: one-dimensional systems, Journal of Computational Physics, 84, 1989, 90-113.

[7] B. Cockburn and C.-W. Shu, TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws II: general framework, Mathematics of Computation, 52, 1989, 411-435.

[8] B. Cockburn and C.-W. Shu, The Runge-Kutta local projection-discontinuous-Galerkin finite element method for scalar conservation laws, ESAIM: Mathematical Modelling and Numerical Analysis, 25, 1991, 337-361.

[9] B. Cockburn and C.-W. Shu, The Runge–Kutta discontinuous Galerkin method for conservation laws V: multidimensional systems, Journal of Computational Physics, 141, 1998, 199-224.

[10] W.A. Fiveland, Discrete-ordinates solutions of the radiative transport equation for rectangular enclosures, Journal of Heat Transfer, 106, 1984, 699–706.

[11] J.L. Guermond, M. Nazarov, B. Popov and I. Tomas, Second-order invariant domain preserving approximation of the Euler equations using convex limiting, SIAM Journal on Scientific Computing, 40, 2018, A3211-A3239.

[12] H. Guo, W. Feng, Z. Xu and Y. Yang, Conservative numerical methods for the reinterpreted discrete fracture model on non-conforming meshes and their applications in contaminant transportation in fractured porous media, Advances in Water Resources, 153, 2021, 103951.

[13] H. Guo and Y. Yang, Bound-preserving discontinuous Galerkin method for compressible miscible displacement in porous media, SIAM Journal on Scientific Computing, 39, 2017, A1969-A1990.

[14] D. Kuzmin, H. Hajduk and A. Rupp, Locally bound-preserving enriched Galerkin methods for the linear advection equation, Computers & Fluids, 205, 2020, 104525.

[15] D. Kuzmin, M. Quezada de Luna, D.I. Ketcheson and J. Grüll, Bound-preserving flux limiting for high-order explicit Runge–Kutta time discretizations of hyperbolic conservation laws, Journal of Scientific Computing, 91, 2022, 21.

[16] K.D. Lathrop and B.G. Carlson, Discrete ordinates angular quadrature of the neutron transport equation, Los Alamos Scientific Laboratory, 1964.

[17] J.E. Lavery, Solution of steady-state one-dimensional conservation laws by mathematical programming, SIAM Journal on Numerical Analysis, 26, 1989, 1081-1089.

[18] P. D. Lax and B. Wendroff, Systems of conservation laws, Communications on Pure and Applied Mathematics, 13, 1960, 217-237.

[19] P. Lesaint and P.A. Raviart, On a finite element method for solving the neutron transport equation, Publications mathématiques et informatique de Rennes, S4, 1974, 1-40.

[20] R. J. LeVeque, Numerical Methods for Conservation Laws, Birkhäuser, 1992.

[21] D. Ling, J. Cheng and C.-W. Shu, Conservative high order positivity-preserving discontinuous Galerkin methods for linear hyperbolic and radiative transfer equations, Journal of Scientific Computing, 77, 2018, 1801-1831.

[22] W. Pazner, Sparse invariant domain preserving discontinuous Galerkin methods with subcell convex limiting, Computer Methods in Applied Mechanics and Engineering, 382, 2021, 113876.

[23] T.E. Peterson, A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation, SIAM Journal on Numerical Analysis, 28, 1991, 133-140.

[24] W.H. Reed and T.R. Hill, Triangular mesh methods for the neutron transport equation, Los Alamos Scientific Laboratory, 1973.

[25] F. Renac, Stationary discrete shock profiles for scalar conservation laws with a discontinuous Galerkin method, SIAM Journal on Numerical Analysis, 53, 2015, 1690-1715.

[26] G.R. Richter, An optimal-order error estimate for the discontinuous Galerkin method, Mathematics of Computation, 50, 1988, 75-88.

[27] A.M. Rueda-Ramírez and G.J. Gassner, A subcell finite volume positivity-preserving limiter for DGSEM discretizations of the Euler equations, arXiv preprint, 2021.

[28] M.A. Taylor, B.A. Wingate and L.P. Bos, A cardinal function algorithm for computing multivariate quadrature points, SIAM Journal on Numerical Analysis, 45, 2007, 193-205.

[29] Y. Xing, X. Zhang and C.-W. Shu, Positivity-preserving high order well-balanced discontinuous Galerkin methods for the shallow water equations. Advances in Water Resources, 33, 2010, 1476-1493.

[30] Z. Xu and C.-W. Shu, High order conservative positivity-preserving discontinuous Galerkin method for stationary hyperbolic equations, Journal of Computational Physics, 466, 2022, 111410.

[31] Z. Xu, Y. Yang and H. Guo, High-order bound-preserving discontinuous Galerkin methods for wormhole propagation on triangular meshes, Journal of Computational Physics, 390, 2019, 323-341.

[32] D. Yuan, J. Cheng and C.-W. Shu, High order positivity-preserving discontinuous Galerkin methods for radiative transfer equations, SIAM Journal on Scientific Computing, 38, 2016, A2987-A3019.

[33] M. Zhang, J. Cheng and J. Qiu, High order positivity-preserving discontinuous Galerkin schemes for radiative transfer equations on triangular meshes, Journal of Computational Physics, 397, 2019, 108811.

[34] X. Zhang, On positivity-preserving high order discontinuous Galerkin schemes for compressible Navier–Stokes equations, Journal of Computational Physics, 328, 2017, 301-343.

[35] X. Zhang and C.-W. Shu, On maximum-principle-satisfying high order schemes for scalar conservation laws, Journal of Computational Physics, 229, 2010, 3091-3120.

[36] X. Zhang and C.-W. Shu, On positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes, Journal of Computational Physics, 229, 2010, 8918-8934.

[37] X. Zhang and C.-W. Shu, Positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations with source terms, Journal of Computational Physics, 230, 2011, 1238-1248.

[38] Y. Zhang, X. Zhang and C.-W. Shu, Maximum-principle-satisfying second order discontinuous Galerkin schemes for convection–diffusion equations on triangular meshes, Journal of Computational Physics, 234, 2013, 295-316.