

Is the Classic Convex Decomposition Optimal for Bound-Preserving Schemes in Multiple Dimensions?

Shumo Cui ^{*}, Shengrong Ding [†], Kailiang Wu [‡]

Abstract

Since proposed in [X. Zhang and C.-W. Shu, *J. Comput. Phys.*, 229: 3091–3120, 2010], the Zhang–Shu framework has attracted extensive attention and motivated many bound-preserving (BP) high-order discontinuous Galerkin and finite volume schemes for various hyperbolic equations. A key ingredient in the framework is the decomposition of the cell averages of the numerical solution into a convex combination of the solution values at certain quadrature points, which helps to rewrite high-order schemes as convex combinations of formally first-order schemes. The classic convex decomposition originally proposed by Zhang and Shu has been widely used over the past decade. It was verified, only for the 1D quadratic and cubic polynomial spaces, that the classic decomposition is optimal in the sense of achieving the mildest BP CFL condition. *Yet, it remained unclear whether the classic decomposition is optimal in multiple dimensions.* In this paper, we find that the classic *multidimensional* decomposition based on the tensor product of Gauss–Lobatto and Gauss quadratures is generally *not* optimal, and we discover a novel alternative decomposition for the 2D and 3D polynomial spaces of total degree up to 2 and 3, respectively, on Cartesian meshes. Our new decomposition allows a larger BP time step-size than the classic one, and moreover, it is rigorously proved to be *optimal* to attain the mildest BP CFL condition, yet requires much fewer nodes. The discovery of such an optimal convex decomposition is highly nontrivial yet meaningful, as it may lead to an improvement of high-order BP schemes for a large class of hyperbolic or convection-dominated equations, at the cost of only a slight and local modification to the implementation code. Several numerical examples are provided to further validate the advantages of using our optimal decomposition over the classic one in terms of efficiency.

^{*}Department of Mathematics and SUSTech International Center for Mathematics, Southern University of Science and Technology, Shenzhen 518055, China (cuism@sustech.edu.cn).

[†]Department of Mathematics, Southern University of Science and Technology, Shenzhen 518055, China (dingsr@sustech.edu.cn).

[‡]Corresponding author. Department of Mathematics and SUSTech International Center for Mathematics, Southern University of Science and Technology, Shenzhen 518055, China; National Center for Applied Mathematics Shenzhen (NCAMS), Shenzhen 518055, China (wukl@sustech.edu.cn). The work of K. Wu. is supported in part by National Natural Science Foundation of China (grant No. 12171227).

1 Introduction

This paper is concerned with high-order robust numerical schemes for hyperbolic conservation laws

$$\begin{cases} u_t + \nabla \cdot \mathbf{f}(u) = 0, & (\mathbf{x}, t) \in \mathbb{R}^d \times \mathbb{R}^+, \\ u(\mathbf{x}, 0) = u_0(\mathbf{x}), & \mathbf{x} \in \mathbb{R}^d, \end{cases} \quad (1)$$

where \mathbf{x} denotes the spatial coordinate variable(s) in d -dimensional space, t denotes the time, the conservative variable(s) u takes values in \mathbb{R}^m , and the flux $\mathbf{f} = (f_1, \dots, f_d)$ takes values in $(\mathbb{R}^m)^d$. Our discussions in this paper can also be applicable to other related hyperbolic or convection-dominated equations.

Solutions to the hyperbolic equations (1) often satisfy certain bounds, which constitute a convex invariant region $G \subset \mathbb{R}^m$. When numerically solving such hyperbolic equations, it is highly desirable or even essential to preserve the intrinsic bounds, namely, to preserve the numerical solutions in the region G . In fact, if the numerical solutions go outside the bounds, for example, negative density or pressure is produced when solving the Euler equations, the discrete problem would become ill-posed due to the loss of hyperbolicity of the system, and may lead to the instability or breakdown of the numerical computation.

As well known, first-order accurate monotone schemes, such as the Godunov scheme, the Lax–Friedrichs scheme, and the Engquist–Osher scheme, are bound-preserving (BP) for scalar conservation laws and many hyperbolic systems. However, seeking high-order accurate BP schemes is rather nontrivial. In the pioneering work of [1, 2], Zhang and Shu proposed a general framework of designing high-order BP discontinuous Galerkin (DG) and finite volume (FV) schemes for hyperbolic conservation laws on rectangular meshes, later generalized to triangular meshes in [3]. Over the past decade, the Zhang–Shu framework has attracted extensive attention and been applied to various hyperbolic systems (e.g., [4, 5, 6, 7, 8, 9, 10, 11, 12]) and convection-dominated equations (e.g., [13, 14, 15, 16, 17]). Recently, motivated by a series of BP works [18, 19, 20, 21] for magnetohydrodynamics, the geometric quasilinearization (GQL) framework was proposed in [22] for studying BP problems involving nonlinear constraints. For more developments on high-order BP schemes, we refer the reader to the review papers [23, 24, 25] and some other BP techniques [26, 27, 28, 29].

An essential ingredient in the Zhang–Shu framework [1, 2] is to decompose the cell averages of the numerical solution into a convex combination of the solution values at certain quadrature points. Based on such a convex decomposition, one can reformulate a one-dimensional (1D) or multidimensional high-order FV or DG scheme into a convex combination of formally 1D first-order schemes. This reformulation then leads to a sufficient condition for the BP property of the updated cell averages, combined with a simple local scaling limiter that can enforce the sufficient condition without losing the high-order accuracy [1, 2].

To illustrate the role of convex decomposition in Zhang–Shu’s BP framework, let us consider a $(k + 1)$ th-order FV or DG scheme for 1D conservation laws with reconstructed or DG polynomials of degree k , denoted by $p_i(x)$, on cell $\Omega_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$. With forward Euler time discretization, the evolution equation of cell averages

can be written as

$$\bar{u}_i^{n+1} = \bar{u}_i^n - \lambda \left(\hat{f}(u_{i+\frac{1}{2}}^-, u_{i+\frac{1}{2}}^+) - \hat{f}(u_{i-\frac{1}{2}}^-, u_{i-\frac{1}{2}}^+) \right), \quad (2)$$

where \bar{u}_i^n is the cell average of $p_i(x)$ on Ω_i at time level n , $\lambda = \Delta t / \Delta x$ is the ratio of the temporal and spatial step-sizes, $u_{i-\frac{1}{2}}^+ = p_i(x_{i-\frac{1}{2}})$ and $u_{i+\frac{1}{2}}^- = p_i(x_{i+\frac{1}{2}})$ for all i . Here, $\hat{f}(\cdot, \cdot)$ is a BP numerical flux with which the first-order scheme is BP under a suitable CFL condition $a\lambda \leq c_0$, where a denotes the maximum characteristic speed, and c_0 is the maximum allowable CFL number for the corresponding first-order scheme. Note that the L -point Gauss–Lobatto quadrature with $L = \lceil \frac{k+3}{2} \rceil$ is exact for polynomials of degree up to k . This implies the following convex decomposition [1]:

$$\bar{u}_i^n = \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} p_i(x) dx = \omega_1^{\text{GL}} u_{i-\frac{1}{2}}^+ + \omega_L^{\text{GL}} u_{i+\frac{1}{2}}^- + \sum_{s=2}^{L-1} \omega_s^{\text{GL}} p_i(x_{i,s}^{\text{GL}}), \quad (3)$$

where $\{\omega_s^{\text{GL}}\}$ are the Gauss–Lobatto weights with $\omega_1^{\text{GL}} = \omega_L^{\text{GL}} = 1/(L(L-1))$, and $\{x_{i,s}^{\text{GL}}\}$ are the quadrature nodes on Ω_i with $x_{i,1}^{\text{GL}} = x_{i-\frac{1}{2}}$ and $x_{i,L}^{\text{GL}} = x_{i+\frac{1}{2}}$. Based on the decomposition (3), Zhang and Shu [1, 2] rewrote the scheme (2) as

$$\bar{u}_i^{n+1} = \omega_1^{\text{GL}} \Pi_1 + \omega_L^{\text{GL}} \Pi_L + \sum_{s=2}^{L-1} \omega_s^{\text{GL}} p_i(x_{i,s}^{\text{GL}}), \quad (4)$$

where

$$\Pi_1 := u_{i-\frac{1}{2}}^+ - \frac{\lambda}{\omega_1^{\text{GL}}} \left(\hat{f}(u_{i-\frac{1}{2}}^+, u_{i+\frac{1}{2}}^-) - \hat{f}(u_{i-\frac{1}{2}}^-, u_{i-\frac{1}{2}}^+) \right), \quad \Pi_L := u_{i+\frac{1}{2}}^- - \frac{\lambda}{\omega_L^{\text{GL}}} \left(\hat{f}(u_{i+\frac{1}{2}}^-, u_{i+\frac{1}{2}}^+) - \hat{f}(u_{i-\frac{1}{2}}^+, u_{i+\frac{1}{2}}^-) \right)$$

are of the same form as the three-point first-order scheme with a scaled time step-size. Thanks to the convex decomposition (4) and the BP property of the first-order scheme, if we use a BP limiter [1] to enforce

$$p_i(x_{i,s}^{\text{GL}}) \in G \quad \forall i, s, \quad (5)$$

then by the convexity of G , the high-order scheme (2) preserves $\bar{u}_j^{n+1} \in G$ under the CFL condition

$$a\lambda \leq c_0 \omega_1^{\text{GL}}. \quad (6)$$

As we have seen, the convex decomposition (3) plays a critical role in constructing high-order BP schemes. It determines not only the theoretical BP CFL condition (6) of the resulting scheme, but also the points (5) to perform the BP limiter. In fact, one may choose a different convex decomposition in the above analysis. In the 1D case, any type of quadrature rule, with weights all positive and nodes including the two endpoints, would give a feasible convex decomposition. However, different decomposition would affect the theoretical BP CFL condition and thus the computational costs. It is natural to ask what decomposition is optimal in the sense of achieving the mildest BP CFL condition. Zhang and Shu mentioned in [1, Remark 2.7] that they checked, for $k = 2, 3$, that their 1D decomposition (3) is optimal. For $k \geq 4$, the optimality of the 1D decomposition (3) has not been proved yet.

This paper aims to make the first attempt at questing the optimal convex decomposition in the *multidimensional* cases. In the multiple dimensions, Zhang and Shu [1, 2] proposed the classic convex decomposition based on the tensor product of the Gauss–Lobatto quadrature and the Gauss quadrature rules; see (13). As the 1D case, their decomposition has been an important foundation for constructing high-order BP multidimensional schemes. Over the past decade, the classic Zhang–Shu decomposition has been widely adopted in designing many high-order BP schemes for various hyperbolic or convection-dominated equations. It is natural to ask the following question:

Is the classic convex decomposition optimal in multiple dimensions?

In this work, we find, in the *multidimensional* cases, that the classic decomposition is generally not optimal for the \mathbb{P}^k spaces, *i.e.* the multivariate polynomial spaces of total degree up to k . *Seeking the optimal convex decomposition in the multidimensional cases is highly complicated and challenging.* In this paper, we restrict our attention to two commonly used spaces (\mathbb{P}^2 and \mathbb{P}^3), which are typically used in the third-order and fourth-order DG schemes, on the Cartesian meshes. For these polynomial spaces, *we discover a novel alternative decomposition, which is rigorously proved to be optimal, namely, to attain the mildest BP CFL condition, yet requires much fewer nodes.* Based on our novel optimal convex decomposition, we can establish more efficient high-order BP DG schemes in the Zhang–Shu framework, as it allows a notably larger BP time step-size than the classic one. The discovery of our optimal convex decomposition is highly nontrivial and may have a broad impact, as it would lead to an overall improvement of third-order and fourth-order BP schemes for a large class of hyperbolic or convection-dominated equations at the cost of only a slight and local modification to the implementation code. We will present several numerical examples to further validate the remarkable advantages of using our optimal decomposition over the classic one in terms of efficiency. It is worth mentioning that seeking the optimal convex decomposition for general \mathbb{P}^k spaces ($k \geq 4$) in the multidimensional cases (on the Cartesian meshes or unstructured meshes) seems challenging and is still open. We hope the present paper could be helpful for motivating further discussions on this interesting problem in the future.

2 General convex decomposition for 2D high-order BP schemes

This section discusses the general feasible convex decomposition for constructing 2D high-order BP DG schemes within the Zhang–Shu framework.

Let \mathbb{P}^k denote the space of multivariate polynomials of total degree up to k . We consider the $(k + 1)$ th-order \mathbb{P}^k -based DG scheme with the forward Euler time discretization for solving the 2D hyperbolic conservation laws

$$u_t + f_1(u)_x + f_2(u)_y = 0, \quad (x, y, t) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+. \quad (7)$$

All our discussions in this paper are also valid for high-order strong-stability-preserving (SSP) time discretizations [30], as they are convex combinations of forward Euler step. Following the Zhang–Shu framework [1, 2],

in order to design a BP DG scheme, we only need to ensure the cell averages within the region G . As long as the BP property of the updated cell averages is guaranteed, one may employ a simple BP limiter to enforce the pointwise bounds of the piecewise DG polynomial solutions without affecting the high-order accuracy [1, 2].

On a rectangular cell $\Omega_{ij} := [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$, the evolution equation of cell averages for the $(k+1)$ th-order DG scheme reads

$$\bar{u}_{ij}^{n+1} = \bar{u}_{ij}^n - \frac{\Delta t}{\Delta x} \sum_{q=1}^Q \omega_q^G \left[\hat{f}_1(u_{i+\frac{1}{2},q}^-, u_{i+\frac{1}{2},q}^+) - \hat{f}_1(u_{i-\frac{1}{2},q}^-, u_{i-\frac{1}{2},q}^+) \right] - \frac{\Delta t}{\Delta y} \sum_{q=1}^Q \omega_q^G \left[\hat{f}_2(u_{q,j+\frac{1}{2}}^-, u_{q,j+\frac{1}{2}}^+) - \hat{f}_2(u_{q,j-\frac{1}{2}}^-, u_{q,j-\frac{1}{2}}^+) \right], \quad (8)$$

where

$$u_{i-\frac{1}{2},q}^+ = p_{ij}(x_{i-\frac{1}{2}}, y_{j,q}^G), \quad u_{i+\frac{1}{2},q}^- = p_{ij}(x_{i+\frac{1}{2}}, y_{j,q}^G), \quad u_{q,j-\frac{1}{2}}^+ = p_{ij}(x_{i,q}^G, y_{j-\frac{1}{2}}), \quad u_{q,j+\frac{1}{2}}^- = p_{ij}(x_{i,q}^G, y_{j+\frac{1}{2}})$$

with $p_{ij}(x, y) \in \mathbb{P}^k$ denoting the DG solution polynomial on Ω_{ij} at time level n satisfying

$$\bar{u}_{ij}^n = \frac{1}{\Delta x \Delta y} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} p_{ij}(x, y) dy dx,$$

and $\{x_{i,q}^G\}_{q=1}^Q$ and $\{y_{j,q}^G\}_{q=1}^Q$ respectively denote the Q -point Gauss quadrature nodes in the intervals $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ and $[y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$, with the corresponding quadrature weights $\{\omega_q^G\}$ satisfying $\sum_{q=1}^Q \omega_q^G = 1$. For the \mathbb{P}^k -based DG scheme, Q is typically taken as $k+1$, such that the quadrature has sufficiently high-order accuracy.

In (8), we take the numerical fluxes \hat{f}_1 and \hat{f}_2 as the BP numerical fluxes with which the corresponding 1D three-point first-order schemes are BP, i.e., for any $u_1, u_2, u_3 \in G$ it holds that

$$u_2 - \frac{\Delta t}{\Delta x} (\hat{f}_1(u_2, u_3) - \hat{f}_1(u_1, u_2)) \in G, \quad u_2 - \frac{\Delta t}{\Delta y} (\hat{f}_2(u_2, u_3) - \hat{f}_2(u_1, u_2)) \in G \quad (9)$$

under a suitable CFL condition $\max\{a_1 \Delta t / \Delta x, a_2 \Delta t / \Delta y\} \leq c_0$, where a_1 and a_2 denote the maximum characteristic speeds in x - and y -directions, and c_0 is the maximum allowable CFL number for the 1D first-order schemes. For example, typically $c_0 = 1$ for the Lax–Friedrichs flux [2, 10], and $c_0 = \frac{1}{2}$ for the HLL and HLLC fluxes [10].

2.1 Feasible convex decomposition in 2D

Similar to the 1D case (3), the BP analysis and design of a 2D scheme (8) also require certain 2D quadrature rule to decompose the cell average \bar{u}_{ij}^n into a convex combination of the values of p_{ij} at some points. A qualified 2D quadrature, which we call *feasible convex decomposition*, should satisfy three requirements, as defined below.

Definition 1 (Feasible convex decomposition in 2D). A 2D convex decomposition

$$\bar{u}_{ij}^n = \sum_{q=1}^Q \omega_q^G \left[\omega_1^- u_{i-\frac{1}{2},q}^+ + \omega_1^+ u_{i+\frac{1}{2},q}^- + \omega_2^- u_{q,j-\frac{1}{2}}^+ + \omega_2^+ u_{q,j+\frac{1}{2}}^- \right] + \sum_{s=1}^S \omega_s p_{ij}(x_{ij}^{(s)}, y_{ij}^{(s)}) \quad (10)$$

is said to be *feasible* for the polynomial space \mathbb{P}^k , if it simultaneously satisfies the following three conditions:

- (i) the convex decomposition holds exactly for all $p \in \mathbb{P}^k$;
- (ii) the weights $\{\omega_1^\pm, \omega_2^\pm, \omega_s\}$ are all positive (their summation equals one);
- (iii) the internal node set $\mathbb{S}_{ij} := \{(x_{ij}^{(s)}, y_{ij}^{(s)})\}_{s=1}^S \subset \Omega_{ij}$.

Based on the tensor product of the L -point Gauss quadrature (with $L = \lceil \frac{k+3}{2} \rceil$) and the Q -point Gauss–Lobatto quadrature, the cell average \bar{u}_{ij}^n can be decomposed into a convex combination of point values of p_{ij} as follows:

$$\begin{aligned} \bar{u}_{ij}^n &= \frac{1}{\Delta x \Delta y} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} p_{ij}(x, y) dx dy = \sum_{s=1}^L \omega_s^{\text{GL}} \left(\frac{1}{\Delta y} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} p_{ij}(x_{i,s}^{\text{GL}}, y) dy \right) = \sum_{s=1}^L \sum_{q=1}^Q \omega_s^{\text{GL}} \omega_q^{\text{G}} p_{ij}(x_{i,s}^{\text{GL}}, y_{j,q}^{\text{G}}) \\ &= \sum_{q=1}^Q \omega_q^{\text{G}} \omega_1^{\text{GL}} \left[u_{i-\frac{1}{2},q}^+ + u_{i+\frac{1}{2},q}^- \right] + \sum_{s=2}^{L-1} \sum_{q=1}^Q \omega_s^{\text{GL}} \omega_q^{\text{G}} p_{ij}(x_{i,s}^{\text{GL}}, y_{j,q}^{\text{G}}). \end{aligned} \quad (11)$$

Similarly, by applying the quadrature rules in a different order, one obtains

$$\bar{u}_{ij}^n = \sum_{q=1}^Q \omega_q^{\text{G}} \omega_1^{\text{GL}} \left[u_{q,j-\frac{1}{2}}^+ + u_{q,j+\frac{1}{2}}^- \right] + \sum_{s=2}^{L-1} \sum_{q=1}^Q \omega_s^{\text{GL}} \omega_q^{\text{G}} p_{ij}(x_{i,q}^{\text{G}}, y_{j,s}^{\text{GL}}). \quad (12)$$

Zhang–Shu classic convex decomposition. In [1, 2], Zhang and Shu proposed the classic convex decomposition by using the tensor-product decomposition formulas (11) and (12):

$$\begin{aligned} \bar{u}_{ij}^n &= \kappa_1 \cdot \bar{u}_{ij}^n + \kappa_2 \cdot \bar{u}_{ij}^n = \kappa_1 \cdot \text{equation (11)} + \kappa_2 \cdot \text{equation (12)} \\ &= \sum_{q=1}^Q \omega_q^{\text{G}} \omega_1^{\text{GL}} \left[\kappa_1 u_{i-\frac{1}{2},q}^+ + \kappa_1 u_{i+\frac{1}{2},q}^- + \kappa_2 u_{q,j-\frac{1}{2}}^+ + \kappa_2 u_{q,j+\frac{1}{2}}^- \right] + \sum_{s=2}^{L-1} \sum_{q=1}^Q \omega_s^{\text{GL}} \omega_q^{\text{G}} \left[\kappa_1 p_{ij}(x_{i,s}^{\text{GL}}, y_{j,q}^{\text{G}}) + \kappa_2 p_{ij}(x_{i,q}^{\text{G}}, y_{j,s}^{\text{GL}}) \right] \end{aligned} \quad (13)$$

with

$$\kappa_1 := \frac{\frac{a_1}{\Delta x}}{\frac{a_1}{\Delta x} + \frac{a_2}{\Delta y}}, \quad \kappa_2 := \frac{\frac{a_2}{\Delta y}}{\frac{a_1}{\Delta x} + \frac{a_2}{\Delta y}}.$$

This classic convex decomposition has been widely used over the past decade.

Jiang–Liu convex decomposition. In [10], Jiang and Liu used a simpler convex decomposition:

$$\begin{aligned} \bar{u}_{ij}^n &= \frac{1}{2} \cdot \bar{u}_{ij}^n + \frac{1}{2} \cdot \bar{u}_{ij}^n = \frac{1}{2} \cdot \text{equation (11)} + \frac{1}{2} \cdot \text{equation (12)} \\ &= \sum_{q=1}^Q \frac{\omega_q^{\text{G}} \omega_1^{\text{GL}}}{2} \left[u_{i-\frac{1}{2},q}^+ + u_{i+\frac{1}{2},q}^- + u_{q,j-\frac{1}{2}}^+ + u_{q,j+\frac{1}{2}}^- \right] + \sum_{s=2}^{L-1} \sum_{q=1}^Q \frac{\omega_s^{\text{GL}} \omega_q^{\text{G}}}{2} \left[p_{ij}(x_{i,s}^{\text{GL}}, y_{j,q}^{\text{G}}) + p_{ij}(x_{i,q}^{\text{G}}, y_{j,s}^{\text{GL}}) \right]. \end{aligned} \quad (14)$$

Remark 1. Both the Zhang–Shu decomposition (13) and the Jiang–Liu decomposition (14) are examples of 2D feasible convex decomposition, and they share the same (classic) internal node set

$$\mathbb{S}_{ij}^{\text{Zhang–Shu}} = \bigcup_{s=2}^{L-1} \bigcup_{q=1}^Q \{(x_{i,s}^{\text{GL}}, y_{j,q}^{\text{G}}), (x_{i,q}^{\text{G}}, y_{j,s}^{\text{GL}})\}. \quad (15)$$

2.2 BP conditions via general convex decomposition

Motivated by [1, 2, 10], this subsection studies the BP conditions for the scheme (8) via the general feasible convex decomposition (10). One can rewrite (10) as

$$\bar{u}_{ij}^n = \sum_{q=1}^Q \omega_q^G \widehat{\omega}_1 \left(u_{i-\frac{1}{2},q}^+ + u_{i+\frac{1}{2},q}^- \right) + \sum_{q=1}^Q \omega_q^G \widehat{\omega}_2 \left(u_{q,j-\frac{1}{2}}^+ + u_{q,j+\frac{1}{2}}^- \right) + \Pi, \quad (16)$$

where $\widehat{\omega}_1 := \min\{\omega_1^-, \omega_1^+\}$, $\widehat{\omega}_2 := \min\{\omega_2^-, \omega_2^+\}$, and

$$\Pi := \sum_{q=1}^Q \omega_q^G \left[(\omega_1^- - \widehat{\omega}_1) u_{i-\frac{1}{2},q}^+ + (\omega_1^+ - \widehat{\omega}_1) u_{i+\frac{1}{2},q}^- + (\omega_2^- - \widehat{\omega}_2) u_{q,j-\frac{1}{2}}^+ + (\omega_2^+ - \widehat{\omega}_2) u_{q,j+\frac{1}{2}}^- \right] + \sum_{s=1}^S \omega_s p_{ij}(x_{ij}^{(s)}, y_{ij}^{(s)}). \quad (17)$$

By using a local scaling limiter [1, 2], one can enforce the DG solution polynomial p_{ij} to satisfy the desired bounds on the boundary nodes:

$$u_{i-\frac{1}{2},q}^+ \in G, \quad u_{i+\frac{1}{2},q}^- \in G, \quad u_{q,j-\frac{1}{2}}^+ \in G, \quad u_{q,j+\frac{1}{2}}^- \in G, \quad q = 1, \dots, Q, \quad \forall i, j, \quad (18)$$

and on the internal nodes:

$$p_{ij}(x_{ij}^{(s)}, y_{ij}^{(s)}) \in G, \quad s = 1, \dots, S, \quad \forall i, j. \quad (19)$$

Noting that (17) expresses Π as a convex combination of the values in (18) and (19), we conclude that $\Pi \in G$, because G is convex. Substituting the decomposition (16) into (8), one can rewrite the scheme (8) as

$$\bar{u}_{ij}^{n+1} = \sum_{q=1}^Q \omega_q^G \widehat{\omega}_1 (H_{i+\frac{1}{2},q}^- + H_{i-\frac{1}{2},q}^+) + \sum_{q=1}^Q \omega_q^G \widehat{\omega}_2 (H_{q,j+\frac{1}{2}}^- + H_{q,j-\frac{1}{2}}^+) + \Pi \quad (20)$$

with

$$\begin{aligned} H_{i+\frac{1}{2},q}^- &= u_{i+\frac{1}{2},q}^- - \frac{\Delta t}{\widehat{\omega}_1 \Delta x} \left(\hat{f}_1(u_{i+\frac{1}{2},q}^-, u_{i+\frac{1}{2},q}^+) - \hat{f}_1(u_{i-\frac{1}{2},q}^+, u_{i+\frac{1}{2},q}^-) \right), \\ H_{i-\frac{1}{2},q}^+ &= u_{i-\frac{1}{2},q}^+ - \frac{\Delta t}{\widehat{\omega}_1 \Delta x} \left(\hat{f}_1(u_{i-\frac{1}{2},q}^+, u_{i+\frac{1}{2},q}^-) - \hat{f}_1(u_{i-\frac{1}{2},q}^-, u_{i-\frac{1}{2},q}^+) \right), \\ H_{q,j+\frac{1}{2}}^- &= u_{q,j+\frac{1}{2}}^- - \frac{\Delta t}{\widehat{\omega}_2 \Delta y} \left(\hat{f}_2(u_{q,j+\frac{1}{2}}^-, u_{q,j+\frac{1}{2}}^+) - \hat{f}_2(u_{q,j-\frac{1}{2}}^+, u_{q,j+\frac{1}{2}}^-) \right), \\ H_{q,j-\frac{1}{2}}^+ &= u_{q,j-\frac{1}{2}}^+ - \frac{\Delta t}{\widehat{\omega}_2 \Delta y} \left(\hat{f}_2(u_{q,j-\frac{1}{2}}^+, u_{q,j+\frac{1}{2}}^-) - \hat{f}_2(u_{q,j-\frac{1}{2}}^-, u_{q,j-\frac{1}{2}}^+) \right), \end{aligned}$$

which are formally 1D three-point first-order schemes (9) that satisfy

$$H_{i+\frac{1}{2},q}^- \in G, \quad H_{i-\frac{1}{2},q}^+ \in G, \quad H_{q,j+\frac{1}{2}}^- \in G, \quad H_{q,j-\frac{1}{2}}^+ \in G,$$

under the CFL type conditions

$$a_1 \Delta t \leq \widehat{\omega}_1 \Delta x, \quad a_2 \Delta t \leq \widehat{\omega}_2 \Delta y. \quad (21)$$

Because (20) is a convex combination form, by the convexity of G we conclude that $\bar{u}_{ij}^{n+1} \in G$ under the conditions (21), which are equivalent to the following BP CFL condition (22). In summary, we arrive at the following theorem.

Theorem 1 (BP via general convex decomposition). *If there is a 2D feasible convex decomposition in the form of (24) and the solution polynomial p_{ij} satisfies (18) and (19) for all i and j , then the high-order scheme (8) preserves $\bar{u}_{ij}^{n+1} \in G$ under the BP CFL condition*

$$\Delta t \leq c_0 \min \left\{ \frac{\omega_1^- \Delta x}{a_1}, \frac{\omega_1^+ \Delta x}{a_1}, \frac{\omega_2^- \Delta y}{a_2}, \frac{\omega_2^+ \Delta y}{a_2} \right\}. \quad (22)$$

As direct consequences of Theorem 1, we have the following two corollaries.

Corollary 1 (BP via Zhang–Shu convex decomposition). *If for all i and j , the solution polynomial p_{ij} satisfies (18) and $p_{ij}(x, y) \in G$ for all $(x, y) \in \mathbb{S}_{ij}^{\text{Zhang–Shu}}$, then the high-order scheme (8) preserves $\bar{u}_{ij}^{n+1} \in G$ under the BP CFL condition*

$$\left(\frac{a_1}{\Delta x} + \frac{a_2}{\Delta y} \right) \Delta t \leq \omega_1^{\text{GL}} c_0 = \frac{c_0}{L(L-1)} \quad \text{with} \quad L = \left\lceil \frac{k+3}{2} \right\rceil. \quad (23)$$

Corollary 2 (BP via Jiang–Liu convex decomposition). *If for all i and j , the solution polynomial p_{ij} satisfies (18) and $p_{ij}(x, y) \in G$ for all $(x, y) \in \mathbb{S}_{ij}^{\text{Zhang–Shu}}$, then the high-order scheme (8) preserves $\bar{u}_{ij}^{n+1} \in G$ under the BP CFL condition*

$$2 \max \left\{ \frac{a_1}{\Delta x}, \frac{a_2}{\Delta y} \right\} \Delta t \leq \omega_1^{\text{GL}} c_0 = \frac{c_0}{L(L-1)} \quad \text{with} \quad L = \left\lceil \frac{k+3}{2} \right\rceil.$$

3 Optimal 2D convex decomposition for \mathbb{P}^2 and \mathbb{P}^3 on rectangular cells

As we have seen, a convex decomposition like (10) plays a critical role in constructing 2D high-order BP schemes: the choice of decomposition, in particular, the corresponding weights $\{\omega_1^-, \omega_1^+, \omega_2^-, \omega_2^+\}$ affect the resulting BP CFL condition (22). While the feasible convex decomposition approaches are not unique, it is natural to seek the *optimal convex decomposition* such that the resulting BP CFL condition (22) is mildest, i.e., it maximizes

$$\min \left\{ \frac{\omega_1^- \Delta x}{a_1}, \frac{\omega_1^+ \Delta x}{a_1}, \frac{\omega_2^- \Delta y}{a_2}, \frac{\omega_2^+ \Delta y}{a_2} \right\}.$$

Seeking such an optimal convex decomposition in 2D is challenging. We find that the classic Zhang–Shu decomposition (13) and the Jiang–Liu decomposition (14) both are generally not optimal for the \mathbb{P}^k spaces. Moreover, we discover the following novel convex decomposition on $\Omega_{ij} := [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$ for \mathbb{P}^2 and \mathbb{P}^3 .

Optimal 2D convex decomposition. For $p_{ij} \in \mathbb{P}^2$ or \mathbb{P}^3 , the cell average \bar{u}_{ij}^n has the following convex decomposition

$$\bar{u}_{ij}^n = \frac{\mu_1}{2} \sum_{q=1}^Q \omega_q^G \left[u_{i-\frac{1}{2}, q}^+ + u_{i+\frac{1}{2}, q}^- \right] + \frac{\mu_2}{2} \sum_{q=1}^Q \omega_q^G \left[u_{q, j-\frac{1}{2}}^+ + u_{q, j+\frac{1}{2}}^- \right] + \omega \sum_s p_{ij}(\hat{x}_s, \hat{y}_s), \quad (24)$$

with the internal nodes

$$\mathbb{S}_{ij}^{\text{optimal}} = \{(\hat{x}_s, \hat{y}_s)\} = \begin{cases} \left(x_i, y_j \pm \frac{\Delta y}{2\sqrt{3}} \sqrt{\frac{\phi_* - \phi_2}{\phi_*}} \right), & \text{if } \phi_1 \geq \phi_2, \\ \left(x_i \pm \frac{\Delta x}{2\sqrt{3}} \sqrt{\frac{\phi_* - \phi_1}{\phi_*}}, y_j \right), & \text{if } \phi_1 < \phi_2, \end{cases} \quad (25)$$

where

$$\phi_1 = \frac{a_1}{\Delta x}, \quad \phi_2 = \frac{a_2}{\Delta y}, \quad \phi_* = \max\{\phi_1, \phi_2\}, \quad \psi = \phi_1 + \phi_2 + 2\phi_*, \quad \mu_1 = \frac{\phi_1}{\psi}, \quad \mu_2 = \frac{\phi_2}{\psi}, \quad \omega = \frac{\phi_*}{\psi}.$$

It can be verified that the proposed 2D convex decomposition (24) is feasible and optimal (see Theorem 2) for both \mathbb{P}^2 and \mathbb{P}^3 . As a direct consequence of Theorem 1, we have following corollary.

Corollary 3 (BP via optimal convex decomposition). *If for all i and j , the solution polynomial p_{ij} satisfies (18) and $p_{ij}(x, y) \in G$ for all $(x, y) \in \mathbb{S}_{ij}^{\text{optimal}}$, then the \mathbb{P}^2 -based or \mathbb{P}^3 -based high-order DG scheme (8) preserves $\bar{u}_{ij}^{n+1} \in G$ under the BP CFL condition*

$$\left[2\frac{a_1}{\Delta x} + 2\frac{a_2}{\Delta y} + 4\max\left\{\frac{a_1}{\Delta x}, \frac{a_2}{\Delta y}\right\} \right] \Delta t \leq c_0. \quad (26)$$

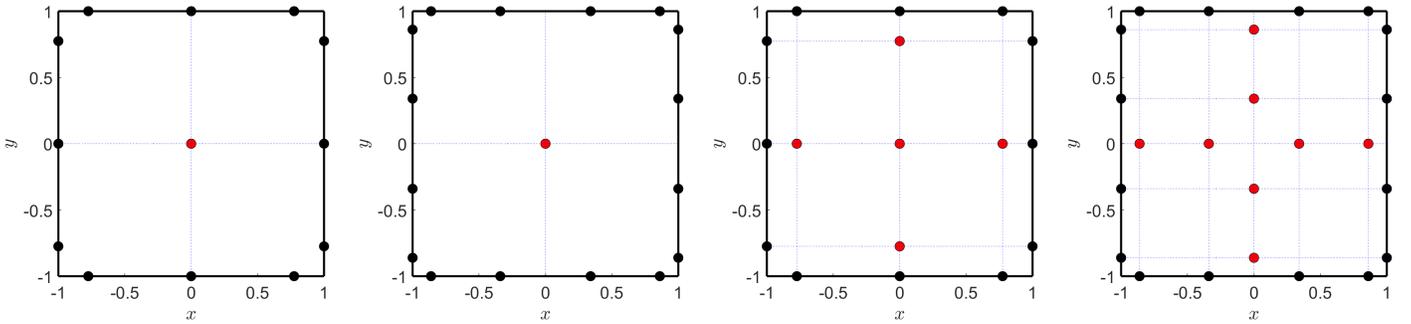
In Table 1, we list and compare the corresponding BP CFL conditions and the internal node sets of decompositions (24), (13) and (14) for \mathbb{P}^2 and \mathbb{P}^3 . We observe that our novel convex decomposition (24) has some remarkable advantages, as summarized in the following remarks.

Remark 2 (Advantage in mildest BP CFL condition). *One can observe from Table 1 that the proposed convex decomposition (24) achieves a notably milder BP CFL condition than the existing ones, i.e., our BP CFL condition (26) is weaker than (13) and (14) respectively obtained via the Zhang–Shu and Jiang–Liu convex decompositions. In fact, for the \mathbb{P}^2 or \mathbb{P}^3 space, no other 2D feasible convex decomposition can achieve an even milder BP CFL condition than (26), i.e., the proposed convex decomposition (24) is optimal. This will be theoretically proved in Theorem 2.*

Remark 3 (Advantage in fewer nodes). *The internal node set $\mathbb{S}_{ij}^{\text{optimal}}$ of the optimal convex decomposition (24) contains only two nodes, which merge to a single node (x_i, y_j) in case of $\phi_1 = \phi_2$. In comparison, the classic convex decomposition (13) or (14) needs much more internal nodes (approximately $2Q(L-2)$ in total). These two internal node sets, $\mathbb{S}_{ij}^{\text{Zhang–Shu}}$ and $\mathbb{S}_{ij}^{\text{optimal}}$, are shown in Figure 1 for further comparison. When the local scaling BP limiter is performed at the internal nodes in all computational cells, using the optimal convex decomposition (24) reduces the computational cost of the BP limiting procedure.*

Table 1: The theoretical BP CFL conditions and the internal node sets of the optimal convex decomposition (24) and the convex decompositions (13) and (14) in 2D for the \mathbb{P}^2 and \mathbb{P}^3 spaces.

	BP CFL condition	BP CFL condition	Internal	Internal
	general case	$\frac{\Delta x}{a_1} = \frac{\Delta y}{a_2} = h$	\mathbb{P}^2 nodes	\mathbb{P}^3 nodes
Optimal	$\left[2\frac{a_1}{\Delta x} + 2\frac{a_2}{\Delta y} + 4 \max \left\{ \frac{a_1}{\Delta x}, \frac{a_2}{\Delta y} \right\} \right] \Delta t \leq c_0$	$\Delta t \leq \frac{c_0}{8} h$	1 ~ 2	1 ~ 2
Zhang & Shu [1]	$\left[6\frac{a_1}{\Delta x} + 6\frac{a_2}{\Delta y} \right] \Delta t \leq c_0$	$\Delta t \leq \frac{c_0}{12} h$	5	8
Jiang & Liu [10]	$\left[12 \max \left\{ \frac{a_1}{\Delta x}, \frac{a_2}{\Delta y} \right\} \right] \Delta t \leq c_0$	$\Delta t \leq \frac{c_0}{12} h$	5	8



(a) Boundary nodes (black) and optimal internal nodes $\mathbb{S}_{ij}^{\text{optimal}}$ (red) for \mathbb{P}^2 . (b) Boundary nodes (black) and optimal internal nodes $\mathbb{S}_{ij}^{\text{optimal}}$ (red) for \mathbb{P}^3 . (c) Boundary nodes (black) and classic internal nodes $\mathbb{S}_{ij}^{\text{Zhang-Shu}}$ (red) for \mathbb{P}^2 . (d) Boundary nodes (black) and classic internal nodes $\mathbb{S}_{ij}^{\text{Zhang-Shu}}$ (red) for \mathbb{P}^3 .

Figure 1: Nodes of the convex decompositions (24) and the classic convex decomposition (13) on $\Omega_{ij} = [-1, 1]^2$, for the \mathbb{P}^2 and \mathbb{P}^3 spaces, in the case of $\frac{\Delta x}{a_1} = \frac{\Delta y}{a_2}$.

Remark 4 (Easy implementation). *It is worth emphasizing that one only requires a slight and local modification to an existing code to enjoy the above-mentioned advantages of our optimal convex decomposition. Specifically, one only needs to replace the classic internal node set $\mathbb{S}_{ij}^{\text{Zhang-Shu}}$ with the optimal internal node set $\mathbb{S}_{ij}^{\text{optimal}}$ in the BP limiting procedure, and then the theoretical BP CFL condition is improved to (26).*

Theorem 2. *For both \mathbb{P}^2 and \mathbb{P}^3 spaces, the 2D convex decomposition (24) is optimal among all feasible candidates.*

Proof. It can be easily verified that the 2D convex decomposition (24) is feasible for \mathbb{P}^2 and \mathbb{P}^3 . We will prove its optimality by contradiction. Assume that there is another feasible convex decomposition of the form

$$\bar{u}_{ij}^n = \sum_{q=1}^Q \omega_q^G \left[\omega_1^- u_{i-\frac{1}{2},q}^+ + \omega_1^+ u_{i+\frac{1}{2},q}^- + \omega_2^- u_{q,j-\frac{1}{2}}^+ + \omega_2^+ u_{q,j+\frac{1}{2}}^- \right] + \sum_{s=1}^S \omega_s p_{ij}(x_{ij}^{(s)}, y_{ij}^{(s)}), \quad (27)$$

which achieves a BP CFL condition milder than (24), that is,

$$c_0 \min \left\{ \frac{\omega_1^- \Delta x}{a_1}, \frac{\omega_1^+ \Delta x}{a_1}, \frac{\omega_2^- \Delta y}{a_2}, \frac{\omega_2^+ \Delta y}{a_2} \right\} > c_0 \min \left\{ \frac{\mu_1 \Delta x}{2a_1}, \frac{\mu_1 \Delta x}{2a_1}, \frac{\mu_2 \Delta y}{2a_2}, \frac{\mu_2 \Delta y}{2a_2} \right\} = \frac{c_0}{2 \frac{a_1}{\Delta x} + 2 \frac{a_2}{\Delta y} + 4 \max \left\{ \frac{a_1}{\Delta x}, \frac{a_2}{\Delta y} \right\}}.$$

In case of $\frac{a_1}{\Delta x} \geq \frac{a_2}{\Delta y}$, we have

$$\begin{aligned} 1 &< \min \left\{ \frac{\omega_1^- \Delta x}{a_1}, \frac{\omega_1^+ \Delta x}{a_1}, \frac{\omega_2^- \Delta y}{a_2}, \frac{\omega_2^+ \Delta y}{a_2} \right\} \left[6 \frac{a_1}{\Delta x} + 2 \frac{a_2}{\Delta y} \right] \\ &\leq 3 \frac{\omega_1^- \Delta x}{a_1} \frac{a_1}{\Delta x} + 3 \frac{\omega_1^+ \Delta x}{a_1} \frac{a_1}{\Delta x} + \frac{\omega_2^- \Delta y}{a_2} \frac{a_2}{\Delta y} + \frac{\omega_2^+ \Delta y}{a_2} \frac{a_2}{\Delta y} = 3(\omega_1^+ + \omega_1^-) + (\omega_2^+ + \omega_2^-). \end{aligned} \quad (28)$$

In case of $\frac{a_1}{\Delta x} < \frac{a_2}{\Delta y}$, we have

$$\begin{aligned} 1 &< \min \left\{ \frac{\omega_1^- \Delta x}{a_1}, \frac{\omega_1^+ \Delta x}{a_1}, \frac{\omega_2^- \Delta y}{a_2}, \frac{\omega_2^+ \Delta y}{a_2} \right\} \left[2 \frac{a_1}{\Delta x} + 6 \frac{a_2}{\Delta y} \right] \\ &\leq \frac{\omega_1^- \Delta x}{a_1} \frac{a_1}{\Delta x} + \frac{\omega_1^+ \Delta x}{a_1} \frac{a_1}{\Delta x} + 3 \frac{\omega_2^- \Delta y}{a_2} \frac{a_2}{\Delta y} + 3 \frac{\omega_2^+ \Delta y}{a_2} \frac{a_2}{\Delta y} = (\omega_1^+ + \omega_1^-) + 3(\omega_2^+ + \omega_2^-). \end{aligned} \quad (29)$$

No matter the hypothetical convex decomposition (27) is feasible for \mathbb{P}^2 or \mathbb{P}^3 , it should hold exactly for $p_{ij}(x, y) = (x - x_i)^2$ and $p_{ij}(x, y) = (y - y_j)^2$. This gives

$$\begin{aligned} \frac{\Delta x^2}{12} &= (\omega_1^+ + \omega_1^-) \frac{\Delta x^2}{4} + (\omega_2^+ + \omega_2^-) \frac{\Delta x^2}{12} + \sum_{s=1}^N \omega_s (x_{ij}^{(s)} - x_i)^2, \\ \frac{\Delta y^2}{12} &= (\omega_1^+ + \omega_1^-) \frac{\Delta y^2}{12} + (\omega_2^+ + \omega_2^-) \frac{\Delta y^2}{4} + \sum_{s=1}^N \omega_s (y_{ij}^{(s)} - y_j)^2, \end{aligned}$$

implying

$$3(\omega_1^+ + \omega_1^-) + (\omega_2^+ + \omega_2^-) \leq 1 \quad \text{and} \quad (\omega_1^+ + \omega_1^-) + 3(\omega_2^+ + \omega_2^-) \leq 1, \quad (30)$$

which contradict with either (28) or (29). Hence the assumption is incorrect, and decomposition (24) is optimal. \square

Remark 5. *The standard CFL condition for linear stability of the \mathbb{P}^k -based DG method with a $(k+1)$ -stage $(k+1)$ -order Runge–Kutta (RK) time discretization [31] is given by the following empirical formula*

$$\left(\frac{a_1}{\Delta x} + \frac{a_2}{\Delta y} \right) \Delta t \leq \frac{1}{2k+1}. \quad (31)$$

Table 2 gives a comparison of different CFL conditions in the special case of $\frac{\Delta x}{a_1} = \frac{\Delta y}{a_2} = h$ for the \mathbb{P}^2 -based (third-order) and \mathbb{P}^3 -based (fourth-order) DG methods. One can see that if $c_0 = 1$, the optimal BP CFL condition (26) of the DG schemes (with the BP limiter) is even weaker than the standard one (31).

Remark 6. *We would like to clarify that the optimal BP CFL condition (26) is merely the best among all those achieved via feasible convex decomposition. It does not mean that such an optimal condition is always sharp or necessary, as other possible analysis approaches may give perhaps weaker BP conditions.*

Table 2: Comparison of different CFL conditions in the special case of $\frac{\Delta x}{a_1} = \frac{\Delta y}{a_2} = h$.

linear stability		$c_0 = 1$		$c_0 = 1/2$	
		optimal	classic	optimal	classic
\mathbb{P}^2	$\Delta t \leq \frac{1}{10}h$	$\Delta t \leq \frac{1}{8}h$	$\Delta t \leq \frac{1}{12}h$	$\Delta t \leq \frac{1}{16}h$	$\Delta t \leq \frac{1}{24}h$
\mathbb{P}^3	$\Delta t \leq \frac{1}{14}h$				

4 Optimal 3D convex decomposition for \mathbb{P}^2 and \mathbb{P}^3 on cuboid cells

The proposed 2D optimal convex decomposition (24) can be extended to 3D and higher dimensions. We denote the maximum characteristic speeds in the x -, y - and z -directions by a_1 , a_2 , and a_3 , respectively. Let $\Omega_{ij\ell} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}] \times [z_{\ell-\frac{1}{2}}, z_{\ell+\frac{1}{2}}]$ be a cuboid cell, with Δx , Δy , and Δz denoting its lengths in the x -, y - and z -directions, respectively.

Optimal 3D convex decomposition. For $p_{ij\ell} \in \mathbb{P}^2$ or \mathbb{P}^3 , the *optimal* 3D convex decomposition on $\Omega_{ij\ell}$ is given by

$$\begin{aligned} \bar{u}_{ij\ell} = & \frac{\mu_1}{2} \sum_{q=1}^Q \sum_{r=1}^Q \omega_q^G \omega_r^G \left[u_{i-\frac{1}{2},q,r}^+ + u_{i+\frac{1}{2},q,r}^- \right] + \frac{\mu_2}{2} \sum_{q=1}^Q \sum_{r=1}^Q \omega_q^G \omega_r^G \left[u_{r,j-\frac{1}{2},q}^+ + u_{r,j+\frac{1}{2},q}^- \right] \\ & + \frac{\mu_3}{2} \sum_{q=1}^Q \sum_{r=1}^Q \omega_q^G \omega_r^G \left[u_{q,r,\ell-\frac{1}{2}}^+ + u_{q,r,\ell+\frac{1}{2}}^- \right] + \frac{\omega}{2} \sum_s p_{ij\ell}(\hat{x}_s, \hat{y}_s, \hat{z}_s), \end{aligned} \quad (32)$$

where $\bar{u}_{ij\ell}$ denotes the cell average of $p_{ij\ell}$ over $\Omega_{ij\ell}$, and

$$u_{i\pm\frac{1}{2},q,r}^\mp = p_{ij\ell}(x_{i\pm\frac{1}{2}}, y_{j,q}^G, z_{\ell,r}^G), \quad u_{r,j\pm\frac{1}{2},q}^\mp = p_{ij\ell}(x_{i,r}^G, y_{j\pm\frac{1}{2}}, z_{\ell,q}^G), \quad u_{q,r,k\pm\frac{1}{2}}^\mp = p_{ij\ell}(x_{i,q}^G, y_{j,r}^G, z_{\ell\pm\frac{1}{2}}).$$

In (32), the weights μ_1 , μ_2 , μ_3 , and ω are given by

$$\mu_1 = \frac{\phi_1}{\psi}, \quad \mu_2 = \frac{\phi_2}{\psi}, \quad \mu_3 = \frac{\phi_3}{\psi}, \quad \omega = \frac{\phi_*}{\psi}$$

with

$$\phi_1 = a_1/\Delta x, \quad \phi_2 = a_2/\Delta y, \quad \phi_3 = a_3/\Delta z, \quad \psi = \phi_1 + \phi_2 + \phi_3 + 2\phi_*, \quad \phi_* = \max\{\phi_1, \phi_2, \phi_3\},$$

and the internal nodes are given by

$$\mathbb{S}_{ij\ell}^{\text{optimal}} = \{(\hat{x}_s, \hat{y}_s, \hat{z}_s)\} = \begin{cases} \left(x_i, y_j \pm \frac{\Delta y}{\sqrt{6}} \sqrt{\frac{\phi_* - \phi_2}{\phi_*}}, z_\ell \right) \text{ and } \left(x_i, y_j, z_\ell \pm \frac{\Delta z}{\sqrt{6}} \sqrt{\frac{\phi_* - \phi_3}{\phi_*}} \right), & \text{if } \phi_1 = \max\{\phi_1, \phi_2, \phi_3\}, \\ \left(x_i, y_j, z_\ell \pm \frac{\Delta z}{\sqrt{6}} \sqrt{\frac{\phi_* - \phi_3}{\phi_*}} \right) \text{ and } \left(x_i \pm \frac{\Delta x}{\sqrt{6}} \sqrt{\frac{\phi_* - \phi_1}{\phi_*}}, y_j, z_\ell \right), & \text{if } \phi_2 = \max\{\phi_1, \phi_2, \phi_3\}, \\ \left(x_i \pm \frac{\Delta x}{\sqrt{6}} \sqrt{\frac{\phi_* - \phi_1}{\phi_*}}, y_j, z_\ell \right) \text{ and } \left(x_i, y_j \pm \frac{\Delta y}{\sqrt{6}} \sqrt{\frac{\phi_* - \phi_2}{\phi_*}}, z_\ell \right), & \text{if } \phi_3 = \max\{\phi_1, \phi_2, \phi_3\}. \end{cases}$$

Theorem 3. *For both \mathbb{P}^2 and \mathbb{P}^3 spaces, the 3D convex decomposition (32) is optimal among all feasible candidates.*

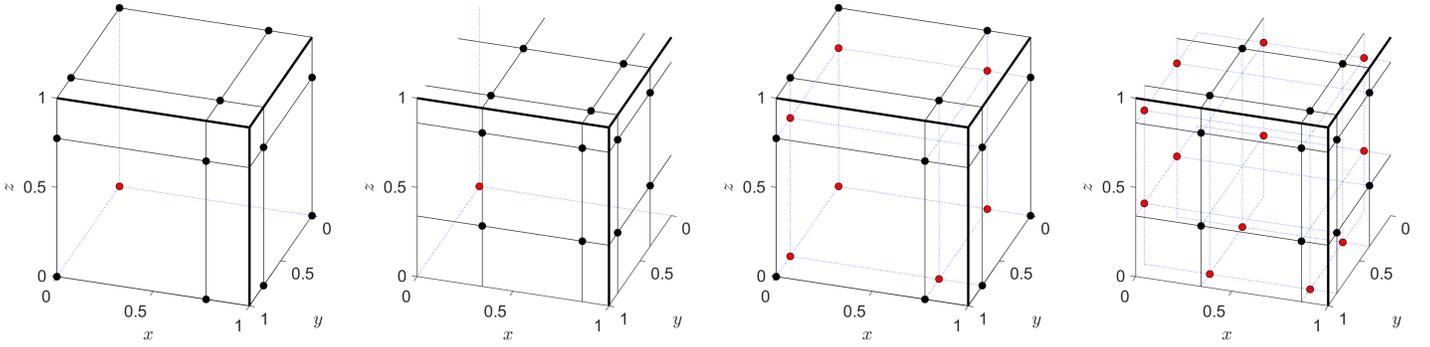
The proof of Theorem 3 is similar to that of Theorem 2 and is thus omitted. As summarized in Table 3, the advantages of the optimal 3D convex decomposition (32), over the 3D versions of the classic decompositions, are even greater than the 2D case. Figure 2 shows the boundary and internal nodes for further comparison.

Table 3: BP CFL conditions and the internal node sets of the optimal 3D convex decomposition (32) and the 3D versions of the Zhang–Shu and Jiang–Liu convex decompositions for the \mathbb{P}^2 and \mathbb{P}^3 spaces.

	BP CFL condition	BP CFL condition	Internal	Internal
	general case	$\frac{\Delta x}{a_1} = \frac{\Delta y}{a_2} = \frac{\Delta z}{a_3} = h$	\mathbb{P}^2 nodes	\mathbb{P}^3 nodes
Optimal	$\left[2\frac{a_1}{\Delta x} + 2\frac{a_2}{\Delta y} + 2\frac{a_3}{\Delta z} + 4 \max\left\{ \frac{a_1}{\Delta x}, \frac{a_2}{\Delta y}, \frac{a_3}{\Delta z} \right\} \right] \Delta t \leq c_0$	$\Delta t \leq \frac{c_0}{10} h$	1 ~ 4	1 ~ 4
Zhang & Shu [1]	$\left[6\frac{a_1}{\Delta x} + 6\frac{a_2}{\Delta y} + 6\frac{a_3}{\Delta z} \right] \Delta t \leq c_0$	$\Delta t \leq \frac{c_0}{18} h$	19	48
Jiang & Liu [10]	$\left[18 \max\left\{ \frac{a_1}{\Delta x}, \frac{a_2}{\Delta y}, \frac{a_3}{\Delta z} \right\} \right] \Delta t \leq c_0$	$\Delta t \leq \frac{c_0}{18} h$	19	48

5 Numerical experiments

In this section, we test the accuracy, robustness, and efficiency of the high-order BP DG schemes designed via the proposed optimal convex decomposition (24), which are referred to as the “optimal approach” for short. For comparison, we also consider the high-order BP DG schemes designed via the classic Zhang–Shu convex decomposition (13), which are referred to as the “classic approach” for short. While the convex decomposition is independent of the choice of BP numerical fluxes, we adopt the global Lax–Friedrichs flux with $c_0 = 1$ in all



(a) Boundary nodes (black) and optimal internal nodes $\mathbb{S}_{ij\ell}^{\text{optimal}}$ (red) for \mathbb{P}^2 . (b) Boundary nodes (black) and optimal internal nodes $\mathbb{S}_{ij\ell}^{\text{optimal}}$ (red) for \mathbb{P}^3 . (c) Boundary nodes (black) and classic internal nodes $\mathbb{S}_{ij\ell}^{\text{Zhang-Shu}}$ (red) for \mathbb{P}^2 . (d) Boundary nodes (black) and classic internal nodes $\mathbb{S}_{ij\ell}^{\text{Zhang-Shu}}$ (red) for \mathbb{P}^3 .

Figure 2: Nodes of the optimal 3D decomposition (32) and the 3D version of the classic Zhang–Shu decomposition on $\Omega_{ij\ell} = [-1, 1]^3$, for \mathbb{P}^2 and \mathbb{P}^3 , in the case of $\frac{\Delta x}{a_1} = \frac{\Delta y}{a_2} = \frac{\Delta z}{a_3}$. We only plot the nodes in the first octant $[0, 1]^3$, while the other nodes are distributed symmetrically.

the presented tests. Unless otherwise stated, we employ the three-stage third-order SSP RK discretization [30] (abbreviated as SSPRK3) for the \mathbb{P}^2 -based (third-order) DG scheme, and the five-stage fourth-order SSP RK time discretization [30] (abbreviated as SSPRK4) for the \mathbb{P}^3 -based (fourth-order) DG scheme. The SSP coefficients for SSPRK3 and SSPRK4 are $C_{\text{SSP}} = 1$ and $C_{\text{SSP}} \approx 1.508$, respectively. All the methods are implemented using C++ language with double precision on a Linux server with Intel(R) Xeon(R) Platinum 8268 CPU @ 2.90GHz 2TB RAM.

Example 1 (Linear convection equation). We start with the two-dimensional linear convection equation

$$u_t + u_x + u_y = 0, \quad (x, y, t) \in [-1, 1] \times [-1, 1] \times \mathbb{R}^+, \quad (33)$$

with periodic boundary conditions and the initial data $u(x, y, 0) = \sin(\pi(x + y))$. The exact solution satisfies a maximum principle, implying the convex invariant region $G = [-1, 1]$. We simulate this problem until $t = 50$ to study the long-time stability of the BP DG schemes with the BP limiter on the uniform mesh of 100×100 cells. Figure 3 shows the time evolution of the numerical errors in the L^1 , L^2 , and L^∞ norms, respectively. In the simulations, we adopt three different time step-sizes, namely, $\tau_0 := C_{\text{SSP}} \tau_0^{\text{BP}}$ with τ_0^{BP} being the maximum Δt determined by the optimal BP CFL condition (26), $\tau_1 := C_{\text{SSP}} \tau_1^{\text{BP}}$ with τ_1^{BP} being the maximum Δt determined by the classic Zhang–Shu BP CFL condition (23), and $\tau_2 := C_{\text{SSP}} \tau_2^{\text{LS}}$ with τ_2^{LS} being the maximum Δt determined by the standard CFL condition (31) for linear stability. The CPU time of all these simulations is presented in Table 4. One can see that all the errors shown in Figure 3 do not exhibit any sign of exponential growth, which indicates the simulations are stable. We also observe that, when the identical time step is used, the errors of the optimal approach is smaller or comparable to those of the classic one, but the optimal approach uses less CPU time (see Table 4) due to the fewer internal nodes. It is seen that the numerical errors are slightly larger, if

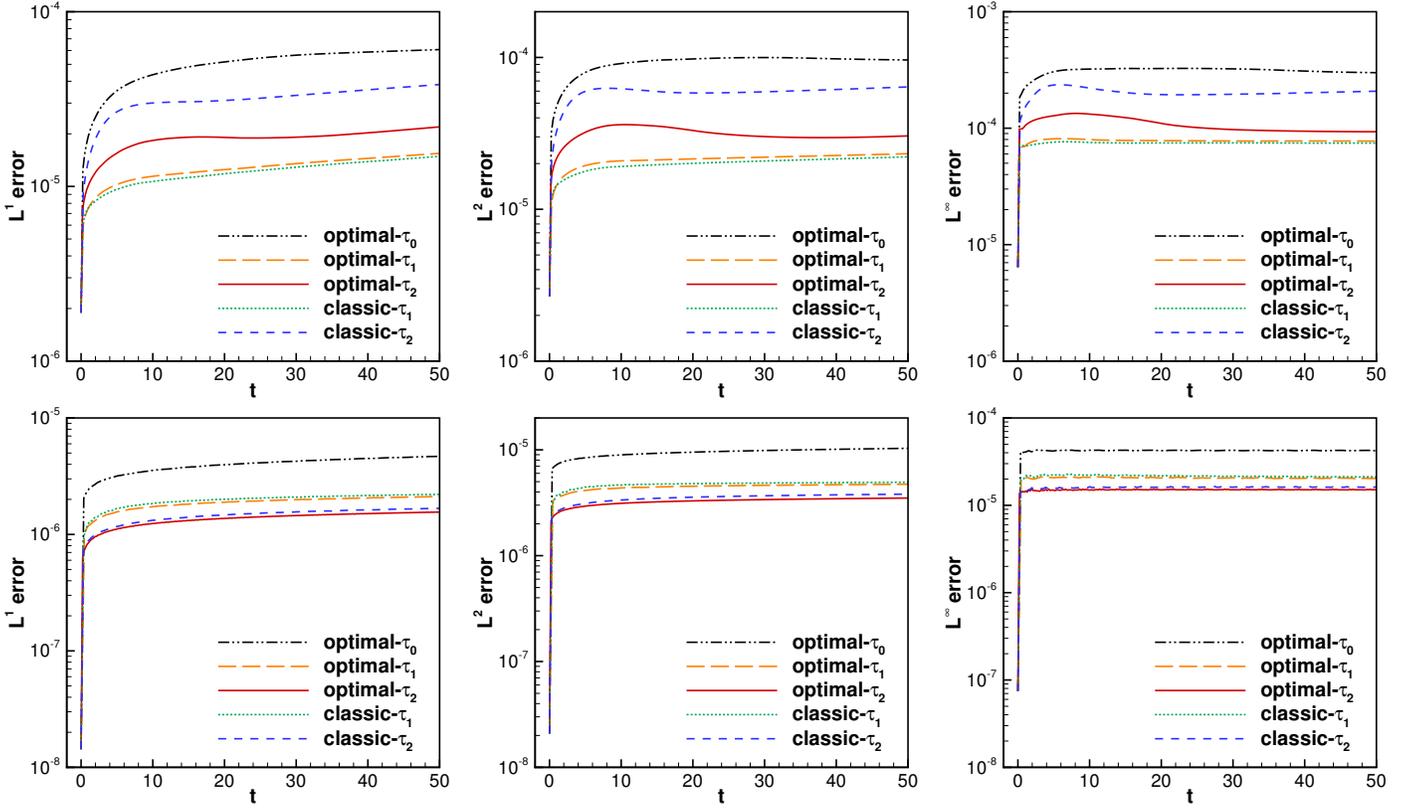


Figure 3: Example 1: Time evolution of the L^1 , L^2 , and L^∞ errors in the numerical solutions obtained using the third-order (\mathbb{P}^2 ; top row) and fourth-order (\mathbb{P}^3 ; bottom row) BP DG schemes designed via different approaches and by using different time step-sizes.

a bigger time step is adopted, as expected. The optimal approach allows a larger time step, with which the CPU time is much less, as shown in Table 4.

Example 2 (Riemann problem of Burgers' equation). This example [32] considers the inviscid Burgers' equation

$$u_t + \left(\frac{u^2}{2}\right)_x + \left(\frac{u^2}{2}\right)_y = 0, \quad (x, y, t) \in [0, 1] \times [0, 1] \times \mathbb{R}^+, \quad (34)$$

with outflow boundary conditions and the initial condition

$$u(x, y, 0) = \begin{cases} -0.2, & \text{if } x < 0.5, y > 0.5, \\ -1, & \text{if } x > 0.5, y > 0.5, \\ 0.5, & \text{if } x < 0.5, y < 0.5, \\ 0.8, & \text{if } x > 0.5, y < 0.5. \end{cases}$$

The exact solution of this example also obeys the maximum principle with $G = [-1, 0.8]$. Figure 4 displays the numerical results at $t = 0.5$ computed with 256×256 uniform cells. We observe that the optimal approach with time step-size $\tau_0 = C_{SSP} \tau_0^{\text{BP}}$ and the classic approach with time step-size $\tau_1 = C_{SSP} \tau_c^{\text{BP}}$ give very similar results,

Table 4: CPU time in seconds for Example 1.

	optimal approach			classic approach	
	τ_0	τ_1	τ_2	τ_1	τ_2
\mathbb{P}^2	1563.22	2410.99	1980.90	2603.67	2166.29
\mathbb{P}^3	3058.73	4564.84	5330.88	4947.13	5775.08

and the discontinuities are equally well resolved by both approaches. However, the CPU time of the optimal approach is much less than that of the classic approach, as exhibited in Table 5.

Table 5: CPU time in seconds for Examples 2–4.

	Approach	Example 2	Example 3	Example 4
		2D Riemann problem	Mach 80 jet	Mach 2000 jet
\mathbb{P}^2	optimal- τ_0	309.39	13297.97	10708.38
	classic- τ_1	504.95	20501.56	16527.88
\mathbb{P}^3	optimal- τ_0	664.71	30861.78	24304.09
	classic- τ_1	1106.00	42714.91	34896.86

Example 3 (Mach 80 jet problem of Euler equations). This example simulates a Mach 80 jet [33, 2, 34] by solving the two-dimensional Euler equations, which can be written in the form of (7) with

$$u = \begin{pmatrix} \rho \\ \rho v_1 \\ \rho v_2 \\ E \end{pmatrix}, \quad f_1(u) = \begin{pmatrix} \rho v_1 \\ \rho v_1^2 + p \\ \rho v_1 v_2 \\ (E + p)v_1 \end{pmatrix}, \quad f_2(u) = \begin{pmatrix} \rho v_2 \\ \rho v_1 v_2 \\ \rho v_2^2 + p \\ (E + p)v_2 \end{pmatrix} \quad (35)$$

with $E = \frac{1}{2}\rho(v_1^2 + v_2^2) + \rho e$ and $p = (\gamma - 1)\rho e$. Here, ρ is the density, (v_1, v_2) denotes the velocity, p is the pressure, E is the total energy, and e is the specific internal energy. The ratio of specific heats γ is set to be $5/3$. The density and pressure should be positive, yielding the invariant region $G = \{u : \rho > 0, p > 0\}$, which is a convex set [2].

Initially, the computation domain $[0, 2] \times [-0.5, 0.5]$ is full of the ambient gas with $(\rho, v_1, v_2, p) = (5, 0, 0, 0.4127)$. The jet with state $(\rho, v_1, v_2, p) = (5, 30, 0, 0.4127)$ is injected into the domain from the left boundary between $y = -0.05$ and 0.05 . All the other boundaries are set as outflow boundary conditions, as in [33, 34]. This is

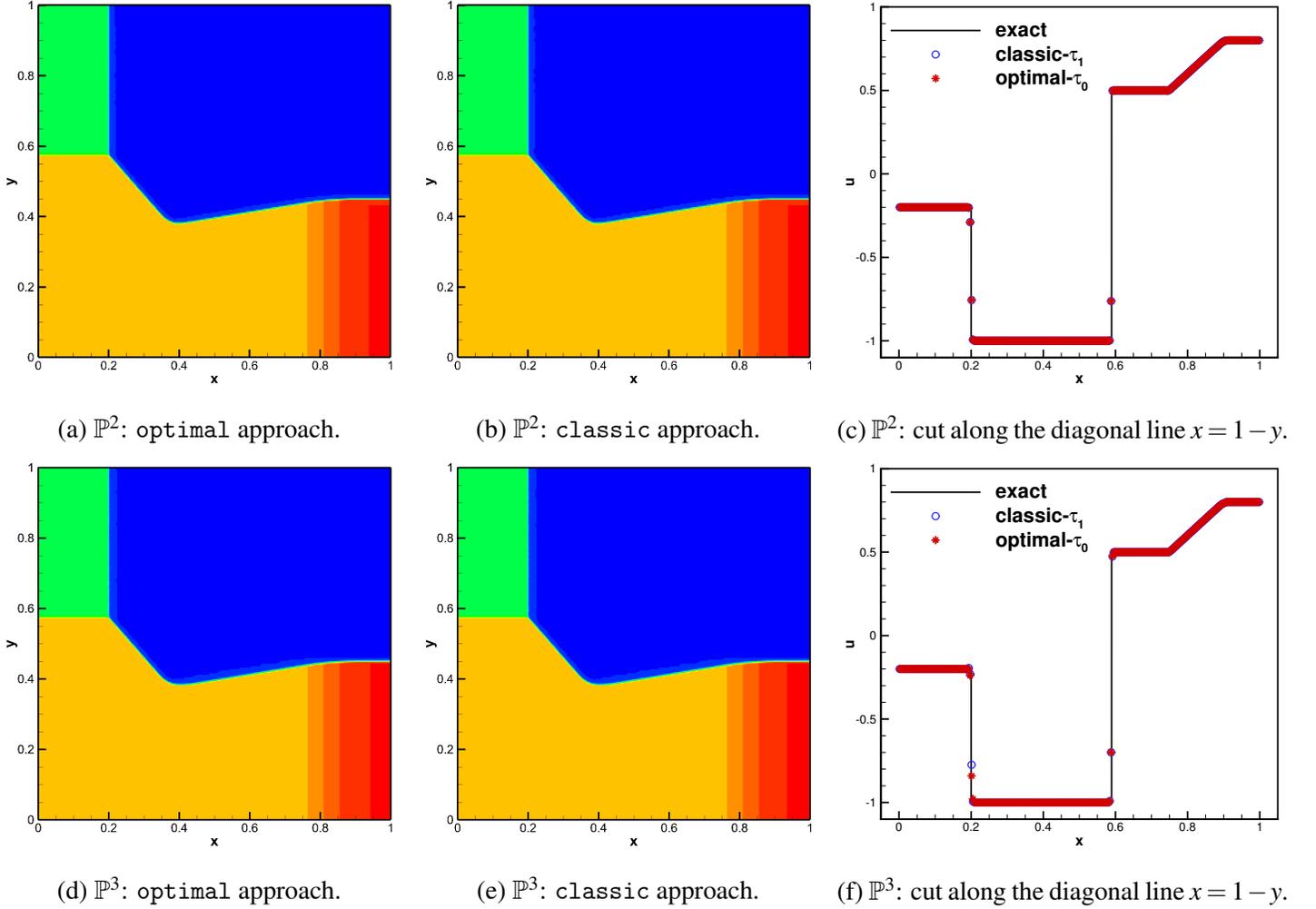


Figure 4: Example 2: The numerical solutions at $t = 0.5$ obtained via the optimal approach and the classic approach.

a benchmark yet challenging test, and a high-order numerical scheme without any BP techniques may easily produce negative density and/or negative pressure, which eventually causes the breakdown of the simulation code. We perform the simulation until $t = 0.07$ on the uniform mesh of 480×240 cells. The numerical results of density are presented in Figure 5, from which we clearly observe that the critical features of the jet: cocoons, bow shock, shear flows, etc. All these features are well captured by the BP DG methods and agree with those presented in [33, 34]. The results of the optimal approach with time step $\tau_0 = C_{SSP} \tau_0^{\text{BP}}$ are comparable to those of the classic approach with time step $\tau_1 = C_{SSP} \tau_1^{\text{BP}}$. As shown in Table 5, the optimal approach allows a larger time step and takes much less CPU time than the classic approach.

For both approaches, the local scaling BP limiter [2] is necessary to enforce the conditions (18) and (19). Due to the presence of strong shocks in this and next examples, the WENO limiter [35] is also used, right before the BP limiter, within some adaptively detected troubled cells to suppress potential numerical oscillations.

Example 4 (Mach 2000 jet problem of Euler equations). Last, a Mach 2000 jet is considered with the Euler

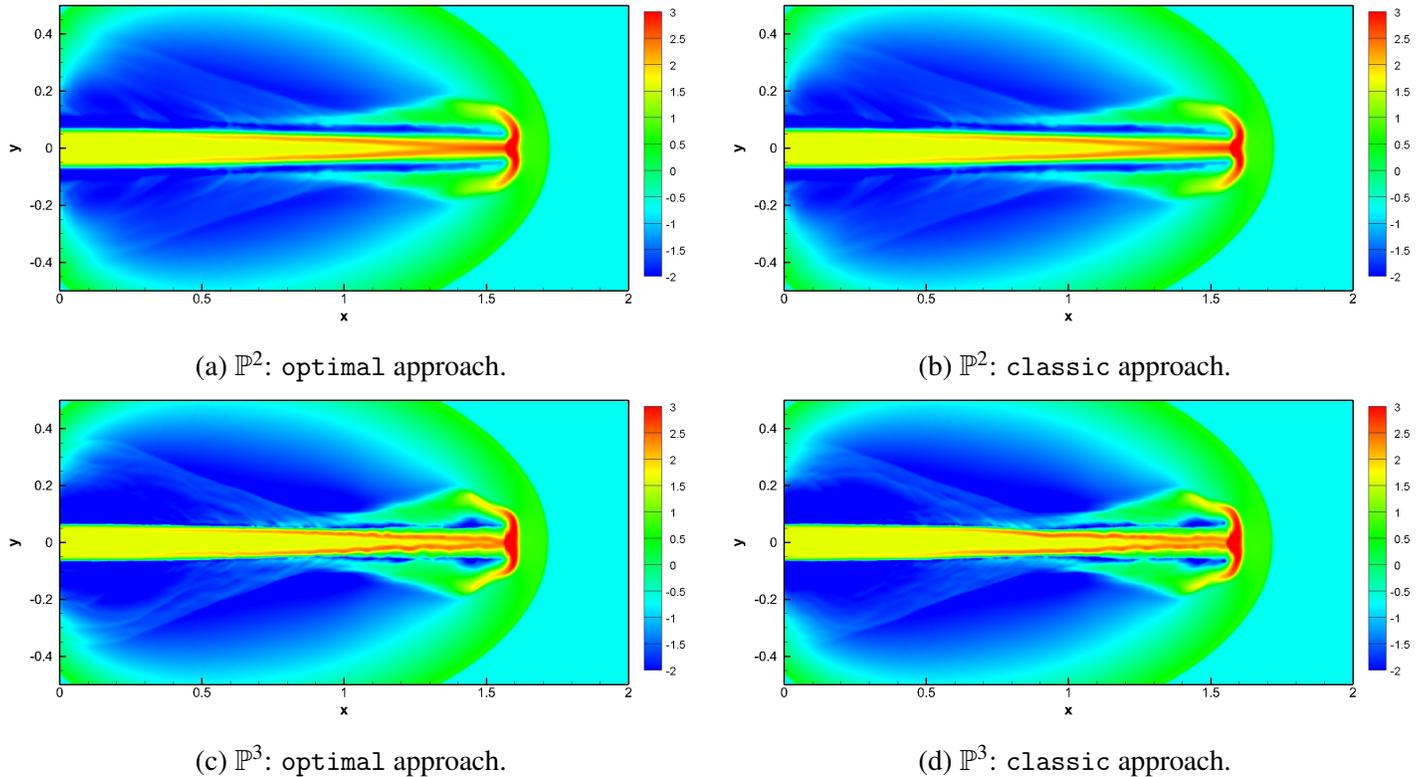


Figure 5: Example 3: The numerical density at $t = 0.07$ obtained by using the third-order (top row) and the fourth-order (bottom row) BP DG schemes designed via the optimal approach and the classic approach.

equation (35) in the domain $[0, 1] \times [-0.25, 0.25]$. The setup is the same as Example 3, except the jet state fixed as $(\rho, v_1, v_2, p) = (5, 800, 0, 0.4127)$ for $y \in [-0.05, 0.05]$ on the left boundary ($x = 0$). All the other boundaries are of outflow conditions. The much higher Mach number renders this jet test more challenging than Example 3. The simulation is performed until $t = 0.001$ with 480×240 cells. Figure 6 presents the numerical results of density, which demonstrate the comparably high resolution and excellent robustness for both the optimal approach and the classic approach. Table 5 also displays the CPU time in this test for both approaches, further confirming the notable advantage of the optimal approach in efficiency.

6 Summary

In this paper, we proposed the problem of seeking the optimal convex decomposition of the cell average for constructing high-order BP schemes of hyperbolic conservation laws in multiple dimensions within the Zhang–Shu framework. It was observed that the classic Zhang–Shu convex decomposition, based on the tensor product of Gauss–Lobatto and Gauss quadratures, is generally not optimal in the multidimensional cases. For the \mathbb{P}^2 and \mathbb{P}^3 spaces, which are typically used in the third-order and fourth-order DG schemes, we discovered the optimal convex decomposition that achieves the mildest BP CFL condition yet requires much fewer internal nodes. Based on our optimal convex decomposition, we established more efficient high-order BP schemes, which allow

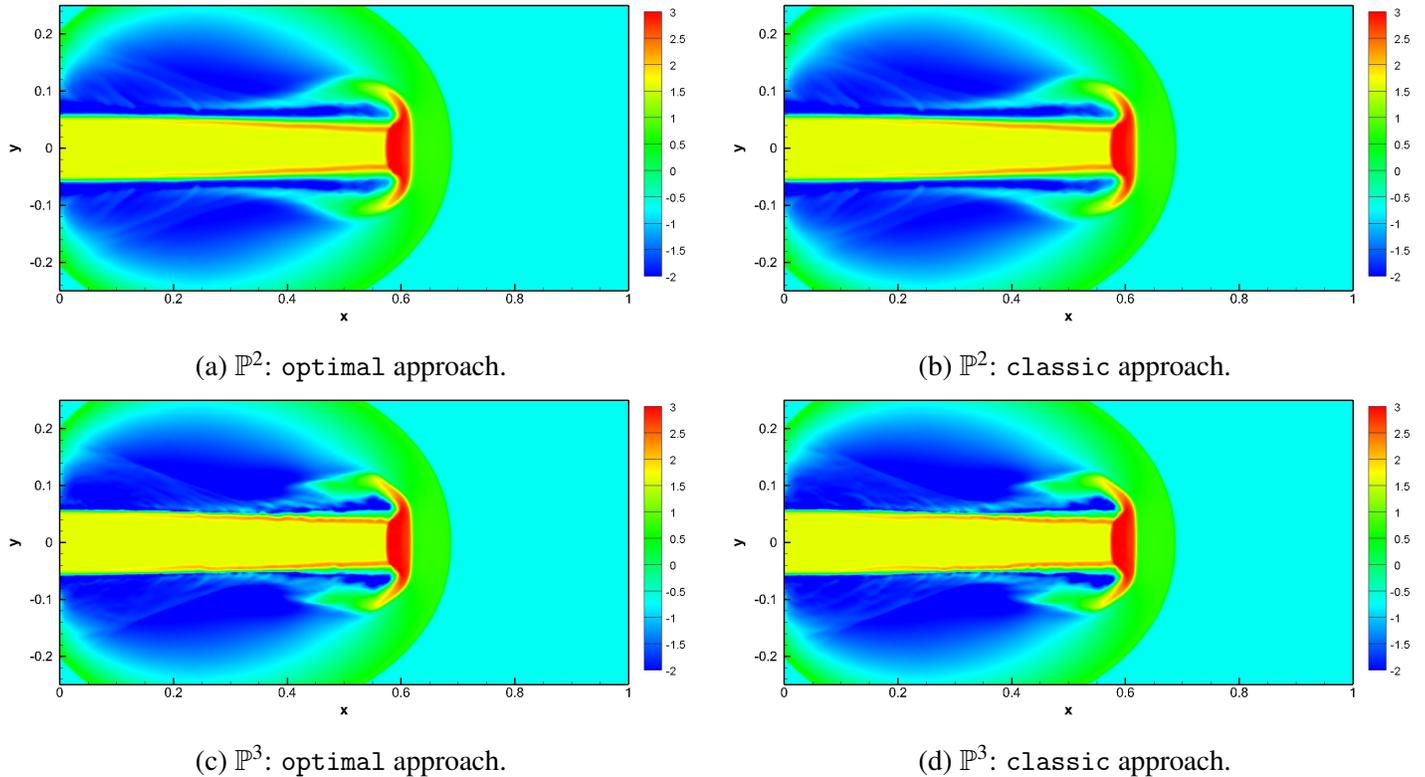


Figure 6: Same as Figure 5 except for Example 4 (Mach 2000 jet) at $t = 0.001$.

a larger BP time step-size than the classic one, in the Zhang–Shu framework. We presented several numerical examples to validate the remarkable advantages of using our optimal decomposition over the classic one in terms of efficiency.

The discovery of the optimal convex decomposition was quite nontrivial and might have a broad impact, as it would lead to an overall improvement of third-order and fourth-order BP schemes for a large class of hyperbolic or convection-dominated equations at the cost of only a slight and local modification to the implementation code. Our work in this paper was limited to the multivariate polynomial spaces \mathbb{P}^2 and \mathbb{P}^3 . In more general cases, many questions about the optimal convex decomposition are yet open; for example, what are the optimal decompositions for more general polynomial spaces \mathbb{P}^k with $k \geq 4$ on Cartesian meshes, triangular meshes, and more general unstructured meshes? We hope this paper could motivate further exploration along this direction in the future.

References

- [1] Xiangxiong Zhang and Chi-Wang Shu. On maximum-principle-satisfying high order schemes for scalar conservation laws. *J. Comput. Phys.*, 229(9):3091–3120, 2010.
- [2] Xiangxiong Zhang and Chi-Wang Shu. On positivity-preserving high order discontinuous Galerkin

- schemes for compressible Euler equations on rectangular meshes. *J. Comput. Phys.*, 229(23):8918–8934, 2010.
- [3] Xiangxiong Zhang, Yinhua Xia, and Chi-Wang Shu. Maximum-principle-satisfying and positivity-preserving high order discontinuous Galerkin schemes for conservation laws on triangular meshes. *J. Sci. Comput.*, 50(1):29–62, 2012.
- [4] Yulong Xing, Xiangxiong Zhang, and Chi-Wang Shu. Positivity-preserving high order well-balanced discontinuous Galerkin methods for the shallow water equations. *Adv. Water Resour.*, 33(12):1476–1493, 2010.
- [5] Xiangxiong Zhang and Chi-Wang Shu. Positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations with source terms. *J. Comput. Phys.*, 230(4):1238–1248, 2011.
- [6] Cheng Wang, Xiangxiong Zhang, Chi-Wang Shu, and Jianguo Ning. Robust high order discontinuous Galerkin schemes for two-dimensional gaseous detonations. *J. Comput. Phys.*, 231(2):653–665, 2012.
- [7] Xiangxiong Zhang and Chi-Wang Shu. A minimum entropy principle of high order schemes for gas dynamics equations. *Numer. Math.*, 121(3):545–563, 2012.
- [8] Tong Qin, Chi-Wang Shu, and Yang Yang. Bound-preserving discontinuous Galerkin methods for relativistic hydrodynamics. *J. Comput. Phys.*, 315:323–347, 2016.
- [9] Kailiang Wu. Design of provably physical-constraint-preserving methods for general relativistic hydrodynamics. *Phys. Rev. D*, 95(10), 2017.
- [10] Yi Jiang and Hailiang Liu. Invariant-region-preserving DG methods for multi-dimensional hyperbolic conservation law systems, with an application to compressible Euler equations. *J. Comput. Phys.*, 373:385–409, 2018.
- [11] Jie Du, Cheng Wang, Chengeng Qian, and Yang Yang. High-order bound-preserving discontinuous Galerkin methods for stiff multispecies detonation. *SIAM J. Sci. Comput.*, 41(2):B250–B273, 2019.
- [12] Kailiang Wu. Minimum principle on specific entropy and high-order accurate invariant region preserving numerical methods for relativistic hydrodynamics. *SIAM J. Sci. Comput.*, 43(6):B1164–B1197, 2021.
- [13] Xiangxiong Zhang, Yuanyuan Liu, and Chi-Wang Shu. Maximum-principle-satisfying high order finite volume weighted essentially nonoscillatory schemes for convection-diffusion equations. *SIAM J. Sci. Comput.*, 34(2):A627–A658, 2012.

- [14] Yifan Zhang, Xiangxiong Zhang, and Chi-Wang Shu. Maximum-principle-satisfying second order discontinuous Galerkin schemes for convection–diffusion equations on triangular meshes. *J. Comput. Phys.*, 234:295–316, 2013.
- [15] Xiangxiong Zhang. On positivity-preserving high order discontinuous Galerkin schemes for compressible Navier-Stokes equations. *J. Comput. Phys.*, 328:301–343, 2017.
- [16] Zheng Sun, José A Carrillo, and Chi-Wang Shu. A discontinuous Galerkin method for nonlinear parabolic equations and gradient flow problems with interaction potentials. *J. Comput. Phys.*, 352:76–104, 2018.
- [17] Jie Du and Yang Yang. Maximum-principle-preserving third-order local discontinuous Galerkin method for convection-diffusion equations on overlapping meshes. *J. Comput. Phys.*, 377:117–141, 2019.
- [18] Kailiang Wu and Huazhong Tang. Admissible states and physical-constraints-preserving schemes for relativistic magnetohydrodynamic equations. *Math. Models Methods Appl. Sci.*, 27(10):1871–1928, 2017.
- [19] Kailiang Wu. Positivity-preserving analysis of numerical schemes for ideal magnetohydrodynamics. *SIAM J. Numer. Anal.*, 56(4):2124–2147, 2018.
- [20] Kailiang Wu and Chi-Wang Shu. A provably positive discontinuous Galerkin method for multidimensional ideal magnetohydrodynamics. *SIAM J. Sci. Comput.*, 40(5):B1302–B1329, 2018.
- [21] Kailiang Wu and Chi-Wang Shu. Provably physical-constraint-preserving discontinuous Galerkin methods for multidimensional relativistic MHD equations. *Numer. Math.*, 148:699–741, 2021.
- [22] Kailiang Wu and Chi-Wang Shu. Geometric quasilinearization framework for analysis and design of bound-preserving schemes. *arXiv preprint arXiv:2111.04722*, 2021.
- [23] Xiangxiong Zhang and Chi-Wang Shu. Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments. *Proc. R. Soc. A*, 467:2752–2776, 2011.
- [24] Zhengfu Xu and Xiangxiong Zhang. Bound-preserving high order schemes. In *Handbook of Numerical Methods for Hyperbolic Problems: Applied and Modern Issues*, edited by R. Abgrall and Chi-Wang Shu, volume 18, pages 81–102, North-Holland, Amsterdam, 2017. Elsevier.
- [25] Chi-Wang Shu. A class of bound-preserving high order schemes: The main ideas and recent developments. In Susanne C. Brenner, Igor E. Shparlinski, Chi-Wang Shu, and Daniel B. Szyld, editors, *75 Years of Mathematics of Computation*, volume 754 of *Contemporary Mathematics*, page 247. American Mathematical Society, 2020.
- [26] Zhengfu Xu. Parametrized maximum principle preserving flux limiters for high order schemes solving hyperbolic conservation laws: one-dimensional scalar problem. *Math. Comp.*, 83(289):2213–2238, 2014.

- [27] Tao Xiong, Jing-Mei Qiu, and Zhengfu Xu. Parametrized positivity preserving flux limiters for the high order finite difference WENO scheme solving compressible Euler equations. *J. Sci. Comput.*, 67(3):1066–1088, 2016.
- [28] Kailiang Wu and Huazhong Tang. High-order accurate physical-constraints-preserving finite difference WENO schemes for special relativistic hydrodynamics. *J. Comput. Phys.*, 298:539–564, 2015.
- [29] Jean-Luc Guermond and Bojan Popov. Invariant domains and second-order continuous finite element approximation for scalar conservation equations. *SIAM J. Numer. Anal.*, 55(6):3120–3146, 2017.
- [30] Sigal Gottlieb, David I Ketcheson, and Chi-Wang Shu. *Strong stability preserving Runge-Kutta and multi-step time discretizations*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2011.
- [31] Bernardo Cockburn and Chi-Wang Shu. Runge–Kutta discontinuous Galerkin methods for convection-dominated problems. *J. Sci. Comput.*, 16(3):173–261, 2001.
- [32] Jianfang Lu, Yong Liu, and Chi-Wang Shu. An oscillation-free discontinuous Galerkin method for scalar hyperbolic conservation laws. *SIAM J. Numer. Anal.*, 59(3):1299–1324, 2021.
- [33] Youngsoo Ha, Carl L. Gardner, Anne Gelb, and Chi-Wang Shu. Numerical simulation of high Mach number astrophysical jets with radiative cooling. *J. Sci. Comput.*, 24(1):29–44, 2005.
- [34] Yong Liu, Jianfang Lu, and Chi-Wang Shu. An essentially oscillation-free discontinuous Galerkin method for hyperbolic systems. *SIAM J. Sci. Comput.*, 44(1):A230–A259, 2022.
- [35] Jianxian Qiu and Chi-Wang Shu. Runge–Kutta discontinuous Galerkin method using WENO limiters. *SIAM J. Sci. Comput.*, 26(3):907–929, 2005.