Bound-preserving high order schemes

Zhengfu Xu[a], Xiangxiong Zhang[b]

[a] *Department of Mathematical Sciences, Michigan Technological University, Houghton, MI 49931*
[b] *Department of Mathematics, Purdue University, West Lafayette, IN 47907*

Abstract

For the initial value problem of scalar conservation laws, a bound-preserving property is desired for numerical schemes in many applications. Traditional methods to enforce a discrete maximum principle by defining the extrema as those of grid point values in finite difference schemes or cell averages in finite volume schemes usually result in an accuracy degeneracy to second order around smooth extrema. On the other hand, successful and popular high order accurate schemes do not satisfy a strict bound-preserving property. We review two approaches for enforcing the bound-preserving property in high order schemes. The first one is a general framework to design a simple and efficient limiter for finite volume and discontinuous Galerkin schemes without destroying high order accuracy. The second one is a bound-preserving flux limiter, which can be used on high order finite difference, finite volume and discontinuous Galerkin schemes.

Keywords: bound-preserving, positivity-preserving, conservation laws, high order accurate schemes, finite difference, finite volume, discontinuous Galerkin

Email addresses: `zhengfux@mtu.edu` (Zhengfu Xu), `zhan1966@purdue.edu` (Xiangxiong Zhang)

1. Introduction

The unique entropy solution $u(\mathbf{x}, t)$ to the initial value problem of the scalar conservation law,

$$u_t + \nabla \cdot \mathbf{F}(u) = 0, \quad u(\mathbf{x}, 0) = u_0(\mathbf{x}), \tag{1}$$

satisfies a strict maximum principle,

$$\min_{\mathbf{x}} u(\mathbf{x}, s) \leq u(\mathbf{x}, t) \leq \max_{\mathbf{x}} u(\mathbf{x}, s), \quad \forall t > s,$$

thus a bound-preserving property $\min_{\mathbf{x}} u_0(\mathbf{x}) \leq u(\mathbf{x}, t) \leq \max_{\mathbf{x}} u_0(\mathbf{x})$. For numerical schemes solving (1), it is desired to have a numerical solution u_j^n satisfying the bound-preserving property

$$\min_{\mathbf{x}} u_0(\mathbf{x}) = m \leq u_j^n \leq M = \max_{\mathbf{x}} u_0(\mathbf{x}), \tag{2}$$

since solutions outside of $[m, M]$ might be meaningless, such as negative percentage and probability distribution larger than one. Moreover, violation of certain bounds in numerical solutions may contribute to instability for systems of equations, e.g., negative density and negative pressure in gas dynamics equations.

The first order accurate E-schemes including Godunov, Lax-Friedrichs and Engquist-Osher methods satisfy an entropy inequality and are total-variation-diminishing (TVD) thus maximum-principle-satisfying. However, any TVD scheme or TVD limiter in the sense of measuring the variation of grid point values or cell averages is at most first order accurate around smooth extrema, see Harten (1983); Osher and Chakravarthy (1984), although these schemes can be designed for any formal order of accuracy for smooth monotone solutions, e.g., the high resolution schemes.

For finite difference and finite volume schemes, a popular approach to achieve bound-preserving is to enforce a discrete maximum principle

$$\min_j u_j^n \leq u_j^{n+1} \leq \max_j u_j^n, \tag{3}$$

where u^n denotes numerical solutions at time step n, and u_j^n denotes the point value of at j-th point in a finite difference scheme or the cell average on j-th cell in a finite volume scheme. Schemes satisfying (3) can be at most second order accurate around extrema, see a simple proof in Zhang and Shu (2011a); Shu (2012). There are many second order accurate schemes satisfying (3) or (2) in the literature, e.g., Bell et al. (1988); Chavent and Cockburn (1989); Colella (1990); Liu (1993); Batten et al. (1996); Jiang and Tadmor (1998); Hubbard (1999); Kurganov and Tadmor (2000); Bouchut (2004); Guermond et al. (2014); Guermond and Nazarov (2014).

One heuristic explanation of why (3) prohibits higher order than second order accuracy is due to the numerical extrema defined as those of grid point values, without including the high order information between two adjacent grid points. Towards higher order accurate schemes, we can measure the extrema in numerical solutions as extrema of approximation polynomials, e.g., reconstruction polynomials in a finite volume scheme. To this end, we can consider an improved discrete maximum principle,

$$\min_j \min_{\mathbf{x}} u_j^n(\mathbf{x}) \leq u_j^{n+1}(\mathbf{x}) \leq \max_j \max_{\mathbf{x}} u_j^n(\mathbf{x}). \tag{4}$$

where $u_j(\mathbf{x})$ denotes the reconstruction polynomial on j-th cell in a finite volume scheme. See Sanders (1988); Liu and Osher (1996); Zhang and Shu (2010a) for third and higher order accurate schemes satisfying (4). However, these schemes used the exact time evolution to enforce (4). Unfortunately, it is very difficult, if not impossible, to implement such exact time evolution for multi-dimensional nonlinear scalar equations or systems of conservation laws.

Successful and popular high order accurate methods solving (1) include, among others, the Runge-Kutta discontinuous Galerkin (RKDG) method with a total variation bounded (TVB) limiter in Cockburn and Shu (1989); Cockburn et al. (1989), the essentially non-oscillatory (ENO) finite volume and finite difference schemes, e.g. Harten et al. (1987); Shu and Osher (1988), and the weighted ENO (WENO) finite volume and finite difference schemes, e.g., Liu et al. (1994); Jiang and Shu (1995). These schemes are nonlinearly stable in numerical experiments and some of them can be proven to be total variation stable, but they are not strictly bound-preserving. To construct high order accurate bound-preserving schemes for multi-dimensional nonlinear equations, instead of enforcing a discrete maximum principle (4), we may consider to enforce only a discrete bound-preserving constraint in a high order scheme. In this paper, we will review two approaches to enforce (2) on these popular high order schemes.

2. A bound-preserving limiter for approximation polynomials

We first review the framework of constructing bound-preserving finite volume and DG schemes in Zhang and Shu (2010b); Zhang et al. (2012); Zhang and Shu (2011a). For simplicity, we only review the main idea for the one-dimensional version of (1):

$$u_t + f(u)_x = 0, \qquad u(x,0) = u_0(x). \tag{5}$$

2.1. First order monotone schemes

Consider a first order scheme with the form

$$u_j^{n+1} = u_j^n - \lambda[\widehat{f}(u_j^n, u_{j+1}^n) - \widehat{f}(u_{j-1}^n, u_j^n)] \equiv H_\lambda(u_{j-1}^n, u_j^n, u_{j+1}^n), \tag{6}$$

where $\lambda = \frac{\Delta t}{\Delta x}$ with Δt and Δx being the temporal and spatial mesh sizes, and $\widehat{f}(a,b)$ is a monotone flux, which is Lipschitz continuous in both arguments, non-decreasing in the first argument and non-increasing in the second argument, and consistent $\widehat{f}(a,a) = f(a)$. Under suitable CFL conditions, typically of the form

$$\max_u |f'(u)|\lambda \leq 1, \tag{7}$$

$H_\lambda(a,b,c)$ is non-increasing in all three arguments, and consistency of $\widehat{f}$ implies $H_\lambda(a,a,a) = a$. For example, consider the Lax-Friedrichs flux $\widehat{f}(u,v) = \frac{1}{2}[f(u) + f(v) - \alpha(v-u)]$ with $\alpha = \max|f'(u)|$, then $H_\lambda(u_{j-1}^n, u_j^n, u_{j+1}^n)$ in the scheme (6) can be rewritten as

$$H_\lambda(u_{j-1}^n, u_j^n, u_{j+1}^n) = (1-\lambda)\, u_j^n + \frac{1}{2}\lambda(\alpha u_{j+1}^n - f(u_{j+1}^n)) + \frac{1}{2}\lambda(\alpha u_{j-1}^n + f(u_{j-1}^n)),$$

which is non-increasing in all three arguments under the constraint (7). Therefore we have the strict maximum principle (3) thus the bound-preserving property

$$m = H_\lambda(m, m, m) \le u_j^{n+1} = H_\lambda(u_{j-1}^n, u_j^n, u_{j+1}^n) \le H_\lambda(M, M, M) = M.$$

2.2. *The weak monotonicity in high order finite volume schemes*

By Godunov's theorem, a linear scheme that is monotone can be at most first order accurate. However, a high order finite volume or DG scheme with a monotone flux for solving (5) satisfies a weak monotonicity property.

We first discuss the forward Euler temporal discretization in this subsection and leave higher order temporal discretization to Section 2.4. The finite volume method or the scheme satisfied by the cell averages in the DG method can be written as:

$$\overline{u}_j^{n+1} = \overline{u}_j^n - \lambda[\widehat{f}(u_{j+\frac{1}{2}}^-, u_{j+\frac{1}{2}}^+) - \widehat{f}(u_{j-\frac{1}{2}}^-, u_{j-\frac{1}{2}}^+)] \equiv G_\lambda(\overline{u}_j^n, u_{j+\frac{1}{2}}^-, u_{j+\frac{1}{2}}^+, u_{j-\frac{1}{2}}^-, u_{j-\frac{1}{2}}^+), \tag{8}$$

where $\overline{u}_j^n$ is the approximation to the cell averages of $u(x,t)$ in the cell $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ at time level n, $\widehat{f}(\cdot,\cdot)$ is a monotone flux, and $u_{j+\frac{1}{2}}^-$, $u_{j+\frac{1}{2}}^+$ are the high order approximations of the nodal values $u(x_{j+\frac{1}{2}}, t^n)$ within the cells I_j and I_{j+1} respectively. For simplicity we assume the mesh is uniform, but the methodology does not have a uniform or smooth mesh restriction.

The function $G_\lambda(a, b, c, d, e)$ in (8) is not a monotonically increasing function for any positive time step. Its partial derivatives with respect to b and e are always non-positive, which implies that there exist $\overline{u}_j^n$, $u_{j+\frac{1}{2}}^-$, $u_{j+\frac{1}{2}}^+$, $u_{j-\frac{1}{2}}^-$, $u_{j-\frac{1}{2}}^+ \in [m, M]$ such that $\overline{u}_j^{n+1} \notin [m, M]$ in (8) for any $\lambda > 0$. In other words, to ensure $\overline{u}_j^{n+1} \in [m, M]$ in (8), we need more constraints on the data $\overline{u}_j^n$, $u_{j+\frac{1}{2}}^-$, $u_{j+\frac{1}{2}}^+$, $u_{j-\frac{1}{2}}^-$, $u_{j-\frac{1}{2}}^+$ other than that they should be in the range $[m, M]$.

Suppose there is a polynomial $p_j(x)$ (either reconstructed in a finite volume method or evolved in a DG method) with degree $k \ge 1$, defined on I_j such that $\overline{u}_j^n$ is the cell average of $p_j(x)$ on I_j, $u_{j-\frac{1}{2}}^+ = p_j(x_{j-\frac{1}{2}})$ and $u_{j+\frac{1}{2}}^- = p_j(x_{j+\frac{1}{2}})$. Let $N = \lfloor (k+3)/2 \rfloor$, i.e., N is smallest integer satisfying $2N - 3 \ge k$. We consider an N-point Legendre Gauss-Lobatto quadrature rule on the interval $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$, which is exact for the integral of polynomials of degree up to $2N-3$. Denote these quadrature points on I_j as

$$S_j = \{x_{j-\frac{1}{2}} = \widehat{x}_j^1, \widehat{x}_j^2, \cdots, \widehat{x}_j^{N-1}, \widehat{x}_j^N = x_{j+\frac{1}{2}}\}. \tag{9}$$

Let $\widehat{\omega}_\mu$ be the N-point Legendre Gauss-Lobatto quadrature weights for the interval $[-\frac{1}{2}, \frac{1}{2}]$ then $\sum\limits_{\mu=1}^{N} \widehat{\omega}_\mu = 1$ and $\widehat{\omega}_1 = \widehat{\omega}_N = \frac{1}{N(N-1)}$. We have

$$\overline{u}_j^n = \frac{1}{\Delta x}\int_{I_j} p_j(x)\ dx = \sum_{\mu=1}^{N} \widehat{\omega}_\mu p_j(\widehat{x}_j^\mu) = \sum_{\mu=2}^{N-1} \widehat{\omega}_\mu p_j(\widehat{x}_j^\mu) + \widehat{\omega}_1 u^+_{j-\frac{1}{2}} + \widehat{\omega}_N u^-_{j+\frac{1}{2}}. \tag{10}$$

After plugging (10) in, the scheme (8) can be rewritten as

$$\begin{aligned}
\overline{u}_j^{n+1} &= \sum_{\mu=2}^{N-1} \widehat{\omega}_\mu p_j(\widehat{x}_j^\mu) + \widehat{\omega}_N \left(u^-_{j+\frac{1}{2}} - \frac{\lambda}{\widehat{\omega}_N}\left[\widehat{f}\left(u^-_{j+\frac{1}{2}}, u^+_{j+\frac{1}{2}}\right) - \widehat{f}\left(u^+_{j-\frac{1}{2}}, u^-_{j+\frac{1}{2}}\right)\right]\right) \\
&\quad + \widehat{\omega}_1 \left(u^+_{j-\frac{1}{2}} - \frac{\lambda}{\widehat{\omega}_1}\left[\widehat{f}\left(u^+_{j-\frac{1}{2}}, u^-_{j+\frac{1}{2}}\right) - \widehat{f}\left(u^-_{j-\frac{1}{2}}, u^+_{j-\frac{1}{2}}\right)\right]\right) \\
&= \sum_{\mu=2}^{N-1} \widehat{\omega}_\mu p_j(\widehat{x}_j^\mu) + \widehat{\omega}_N H_{\lambda/\widehat{\omega}_N}(u^+_{j-\frac{1}{2}}, u^-_{j+\frac{1}{2}}, u^+_{j+\frac{1}{2}}) + \widehat{\omega}_1 H_{\lambda/\widehat{\omega}_1}(u^-_{j-\frac{1}{2}}, u^+_{j-\frac{1}{2}}, u^-_{j+\frac{1}{2}}).
\end{aligned} \tag{11}$$

Let $\widehat{\omega} = \widehat{\omega}_1 = \widehat{\omega}_N$, then $H_{\lambda/\widehat{\omega}}$ is monotone under the CFL condition

$$\max_u |f'(u)|\lambda \le \widehat{\omega} = \frac{1}{N(N-1)}, \tag{12}$$

thus $\overline{u}_j^{n+1}$ in (11) is a monotonically increasing function of all the arguments involved, namely $u^-_{j-\frac{1}{2}}$, $u^+_{j+\frac{1}{2}}$ and $p_j(\widehat{x}_j^\mu)$ for $1 \le \mu \le N$, which is the *weak monotonicity* property for a high order spatial discretization (8). We have the following result on bound-preserving,

Theorem 1. *Consider a finite volume scheme or the scheme satisfied by the cell averages of the DG method (8), associated with the approximation polynomials $p_j(x)$ of degree k (either reconstruction or DG polynomials) in the sense that $\overline{u}_j^n = \frac{1}{\Delta x}\int_{I_j} p_j(x)dx$, $u^+_{j-\frac{1}{2}} = p_j(x_{j-\frac{1}{2}})$ and $u^-_{j+\frac{1}{2}} = p_j(x_{j+\frac{1}{2}})$. A sufficient* *condition for* $\overline{u}_j^{n+1} \in [m, M]$ *is*

$$u^{\pm}_{j-\frac{1}{2}}, u^{\pm}_{j+\frac{1}{2}}, p_j(\widehat{x}_j^{\mu}) \in [m, M], \mu = 2, \cdots, N-1, \tag{13}$$

under the CFL condition (12)*.*

Remark 1. The sufficient condition (13) involves point values $p_j(\widehat{x}_j^{\mu})$ for $\mu = 2, \cdots, N-1$, which are not available in typical ENO and WENO finite volume reconstructions since the polynomial $p_j(x)$ is not reconstructed in ENO and WENO. We can use interpolation to construct an approximation polynomial $p_j(x)$ as in Zhang and Shu (2010b). A better method is to avoid explicitly using point values $p_j(\widehat{x}_j^{\mu})$ for $\mu = 2, \cdots, N-1$. Since $\sum_{\mu=2}^{N-1} \frac{\widehat{\omega}_\mu}{1-2\widehat{\omega}} p_j(\widehat{x}_j^{\mu})$ is a convex combination of point values $p_j(\widehat{x}_j^{\mu})$ for $\mu = 2, \cdots, N-1$, by the Mean Value Theorem, there exists some point $x_j^* \in I_j$ such that

$$\sum_{\mu=2}^{N-1} \frac{\widehat{\omega}_\mu}{1-2\widehat{\omega}} p_j(\widehat{x}_j^{\mu}) = p_j(x_j^*).$$

We can rewrite (11) as

$$\overline{u}_j^{n+1} = (1-2\widehat{\omega}) p_j(x_j^*) + \widehat{\omega} H_{\lambda/\widehat{\omega}}(u^+_{j-\frac{1}{2}}, u^-_{j+\frac{1}{2}}, u^+_{j+\frac{1}{2}}) + \widehat{\omega} H_{\lambda/\widehat{\omega}}(u^-_{j-\frac{1}{2}}, u^+_{j-\frac{1}{2}}, u^-_{j+\frac{1}{2}}),$$

thus we can use the following weaker sufficient condition to replace (13),

$$u^{\pm}_{j-\frac{1}{2}}, u^{\pm}_{j+\frac{1}{2}}, p_j(x_j^*) \in [m, M], \tag{14}$$

where $p_j(x_j^*)$ can be computed as $p_j(x_j^*) = \frac{1}{1-2\widehat{\omega}}(\overline{u}_j^n - \widehat{\omega}_1 u^+_{j-\frac{1}{2}} - \widehat{\omega}_N u^-_{j+\frac{1}{2}})$ by (10), even though the location x_j^* is unknown. See Zhang and Shu (2011a); Cai et al. (2015); Zhang (2016)

2.3. A simple and efficient scaling limiter

The weak monotonicity and Theorem 1 suggest that it is possible to render a high order conservative finite volume or DG scheme bound-preserving if we can control certain point values. To enforce the sufficient condition (13) or (14), we first consider the simple scaling limiter introduced in Liu and Osher (1996). Given piecewise polynomials $p_j(x)$ on each interval $I_j =$

$[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ approximating a smooth function $u(x) \in [m, M]$, with the cell averages $\overline{p}_j = \frac{1}{\Delta x}\int_{I_j} p_j(x)\,dx \in [m, M]$, we seek a modified approximation polynomial $\widetilde{p}_j(x)$ satisfying $\widetilde{p}_j(x) \in [m, M]$ for any $x \in I_j$, with the same cell average $\frac{1}{\Delta x}\int_{I_j} \widetilde{p}_j(x)\,dx = \frac{1}{\Delta x}\int_{I_j} p_j(x)\,dx$. For instance, if $p_j(x)$ is the L^2 projection of $u(x) \in [m, M]$ onto the vector space of polynomials of degree k on the interval I_j, then we have $\overline{p}_j \in [m, M]$ but not necessarily $p_j(x) \in [m, M]$ for any $x \in I_j$.

The following limiter was first discussed in Liu and Osher (1996):

$$\widetilde{p}_j(x) = \theta\left[p_j(x) - \overline{p}_j\right] + \overline{p}_j, \quad \theta = \min\left\{1, \left|\frac{M - \overline{p}_j}{M_j - \overline{p}_j}\right|, \left|\frac{m - \overline{p}_j}{m_j - \overline{p}_j}\right|\right\}, \tag{15a}$$

$$M_j = \max_{x \in I_j} p_j(x), m_j = \min_{x \in I_j} p_j(x). \tag{15b}$$

It is obvious that $\widetilde{p}_j(x) \in [m, M]$ for any $x \in I_j$ and the cell average of $\widetilde{p}_j(x)$ is still $\overline{p}_j$. Moreover, this simple limiter does not destroy the approximation accuracy of $p_j(x)$.

Theorem 2. *For the modified polynomial of degree k in the limiter (15), we have $|p_j(x) - \widetilde{p}_j(x)| \leq C_k \max_{x \in I_j} |p_j(x) - u(x)|$, where C_k is a constant depending only on the polynomial degree k.*

PROOF. We only need to discuss the case that $p_j(x)$ is not a constant and $\theta = \left|\frac{M - \overline{p}_j}{M_j - \overline{p}_j}\right|$. The other cases are similar. Since $\overline{p}_j \leq M$ and $\overline{p}_j \leq M_j$, we have $\theta = (M - \overline{p}_j)/(M_j - \overline{p}_j)$. Therefore,

$$\begin{aligned}
\widetilde{p}_j(x) - p_j(x) &= \theta[p_j(x) - \overline{p}_j] + \overline{p}_j - p_j(x) \\
&= (\theta - 1)[p_j(x) - \overline{p}_j] \\
&= \frac{M - M_j}{M_j - \overline{p}_j}[p_j(x) - \overline{p}_j] \\
&= (M - M_j)\frac{p_j(x) - \overline{p}_j}{M_j - \overline{p}_j}.
\end{aligned}$$

Thus $|\widetilde{p}_j(x) - p_j(x)| \leq |M - M_j| \left|\frac{p_j(x) - \overline{p}_j}{M_j - \overline{p}_j}\right|$. The assumption $\theta = \left|\frac{M - \overline{p}_j}{M_j - \overline{p}_j}\right|$ implies the overshoot $M_j > M$. Suppose $p_j(x^{**}) = M_j$ for some $x^{**} \in I_j$, then $u(x^{**}) \leq M < M_j = p_j(x^{**})$. Thus we have $|M - M_j| \leq |u(x^{**}) - p_j(x^{**})| \leq$

$\max_{x\in I_j}|p_j(x)-u(x)|$. We only need to show $\left|\frac{p_j(x)-\overline{p}_j}{M_j-\overline{p}_j}\right| \leq C_k$. Consider a new polynomial $q(x) = p_j\left(x\Delta x + x_{j-\frac{1}{2}}\right) - \overline{p}_j$. Then $\overline{q} = \int_0^1 q(x)\,dx = 0$, $\max\limits_{x\in[0,1]} q(x) = \max\limits_{x\in I_j} p_j(x) - \overline{p}_j$ and $\min\limits_{x\in[0,1]} q(x) = \min\limits_{x\in I_j} p_j(x) - \overline{p}_j$. We have

$$\left|\frac{p_j(x)-\overline{p}_j}{M_j-\overline{p}_j}\right| = \frac{|q(x)|}{\max\limits_{x\in[0,1]} q(x)} \leq \frac{\max\limits_{x\in[0,1]}|q(x)|}{\max\limits_{x\in[0,1]} q(x)} = \max\left\{\frac{\max\limits_{x\in[0,1]} q(x)}{\max\limits_{x\in[0,1]} q(x)}, \frac{-\min\limits_{x\in[0,1]} q(x)}{\max\limits_{x\in[0,1]} q(x)}\right\}.$$

Thus we only need to prove $\frac{\max\limits_{x\in[0,1]}|q(x)|}{\max\limits_{x\in[0,1]} q(x)} \leq C_k$ or $\left|\frac{\min\limits_{x\in[0,1]} q(x)}{\max\limits_{x\in[0,1]} q(x)}\right| \leq C_k$. For quadratic polynomials $k=2$ in one dimension, $C_k = 3$ was proven by explicit calculations in Liu and Osher (1996). For general k and higher dimensions, there are two different kinds of proof. The first proof is similar to proving the equivalence of two norms in a finite-dimensional Banach space.

Lemma 3. *Let $q(x)$ be a non-constant polynomial of degree k with $\int_0^1 q(x)\,dx = 0$, then*

$$\frac{\max\limits_{x\in[0,1]}|q(x)|}{\max\limits_{x\in[0,1]} q(x)} \leq C_k,$$

where C_k is a constant depending only on k.

Proof. Let V denote the finite dimensional vector space consisting of all polynomials of degree k whose averages on the interval $[0,1]$ are zero. For any $q(x)\in V$, define three functional on V by $f_1[q] = \left|\max\limits_{x\in[0,1]} q(x)\right| = \max\limits_{x\in[0,1]} q(x)$, $f_2[q] = \left|\min\limits_{x\in[0,1]} q(x)\right| = -\min\limits_{x\in[0,1]} q(x)$ and $f_0[q] = \max\limits_{x\in[0,1]}|q(x)| = \max\{f_1[q], f_2[q]\}$. Let $\mathbf{e}_i$ $(i=1,\cdots,k)$ be a basis of V. For any vector $c = \begin{bmatrix} c_1 & \cdots & c_k\end{bmatrix}^T \in \mathbb{R}^k$, define $f^j(c) = f_j\left[\sum\limits_i c_i\mathbf{e}_i\right]$ for $j=0,1,2$. Notice that $f_0[\cdot]$ is a norm of V and can be denoted as $f_0[q] = \|q\|_\infty$ on the interval $[0,1]$.

For any $p(x), q(x)\in V$, f_1 satisfies the following properties (similar ones hold for f_2):

1. $\forall a>0$, $f_1[aq(x)] = \max\limits_{x\in[0,1]} aq(x) = af_1[q(x)]$.
2. $f_1[-q] = \left|\max\limits_{x\in[0,1]} -q(x)\right| = \max\limits_{x\in[0,1]} -q(x) = -\min\limits_{x\in[0,1]} q(x) = f_2[q]$.

3. $f_1[p+q] = \max\limits_{x\in[0,1]}(p+q) \le \max\limits_{x\in[0,1]} p + \max\limits_{x\in[0,1]} q = f_1[p] + f_1[q]$.
4. $f_1[q] = 0 \Rightarrow q \equiv 0$.

Thus, for any $c, d \in \mathbb{R}^k$, we have

$$f^1(c) \le f^1(d) + f^1(c-d) \le f^1(d) + f^0(c-d),$$

and

$$f^1(c) \ge f^1(d) - f^1(d-c) = f^1(d) - f^2(c-d) \ge f^1(d) - f^0(c-d),$$

which implies

$$|f^1(c){-}f^1(d)| \le f^0(c{-}d) = f_0\left[\sum_i (c_i - d_i)\mathbf{e}_i\right] \le \sum_i |c_i{-}d_i| \|\mathbf{e}_i\|_\infty \le \sqrt{\sum_i |c_i - d_i|^2}\sqrt{\|\mathbf{e}_i\|_\infty^2}.$$

Therefore, $f^1(c)$ is uniformly continuous w.r.t. the variable c. Notice that the unit sphere $S^1 = \{c \in \mathbf{R}^k : \|c\| = 1\}$ is a compact set, so f^1 attains its maximum and minimum values on S^1:

$$D_1 \le f^1(d) \le D_2, \quad \forall d \in S^1,$$

where D_1 and D_2 are constants. If there exists $d \in S^1$ such that $f^1(d) = 0$, then $d = \mathbf{0}$ by Property 4 above, which is a contraction to $d \in S^1$. So we have $D_1 > 0$. By Property 1, we get $f^1(c/\|c\|) = f^1(c)/\|c\|$, thus we have

$$0 < D_1\|c\| \le f^1(c) \le D_2\|c\|, \quad \forall c \in \mathbb{R}^k, c \ne \mathbf{0}.$$

Notice that $f^0(c)$ is a norm of $\mathbb{R}^k$, thus by the equivalence of any two norms of $\mathbb{R}^k$, we get

$$0 < D_3\|c\| \le f^0(c) \le D_4\|c\|, \quad \forall c \in \mathbb{R}^k, c \ne \mathbf{0}.$$

Therefore, for $q = \sum_i c_i\mathbf{e}_i$, we have

$$\frac{\max\limits_{x\in[0,1]} |q(x)|}{\max\limits_{x\in[0,1]} q(x)} = \frac{f_0[q]}{f_1[q]} = \frac{f^0(c)}{f^1(c)} \le \frac{D_4}{D_1}.$$

Lemma 4. *Let $q(x)$ be a non-constant polynomial of degree k with $\int_0^1 q(x)\,dx = 0$, then*

$$\left| \frac{\max\limits_{x\in[0,1]} q(x)}{\min\limits_{x\in[0,1]} q(x)} \right| \le (k^2 + k - 1)\Lambda_{k+1}[0,1],$$

where $\Lambda_{k+1}[0,1] = \max\limits_{x\in[0,1]} \sum\limits_{j=1}^{k+1} |l_j(x)|$ is the Lebesgue constant with $l_j(x)$ ($j = 1, \cdots, k+1$) being the Lagrange interpolation polynomials at the $(k+1)$-point Gauss-Lobatto quadrature points on the interval $[0,1]$.

Proof. Let $M' = \max\limits_{x\in[0,1]} q(x)$ and $m' = \min\limits_{x\in[0,1]} q(x)$ then $M' > 0$ and $m' < 0$. If $M' \le -m'$, then $\left|\frac{M'}{m'}\right| \le 1$. We only need to discuss the case $M' > -m'$.

Let $\bar{x}_j$ ($j = 1, \cdots, k+1$) denote the $(k+1)$-point Gauss-Lobatto quadrature points for the interval $[0,1]$ and $\bar{\omega}_j$ ($j = 1, \cdots, k+1$) denote the corresponding weights. Then this quadrature is exact for integration of polynomials of degree k. Since $l_j(x)$ ($j = 1, \cdots, k+1$) are the Lagrange interpolation polynomials, we have

$$q(x) = \sum_{j=1}^{k+1} q(\bar{x}_j) l_j(x).$$

Let $M'' = \max_j q(\bar{x}_j)$ and $m'' = \min_j q(\bar{x}_j)$. If $q(\bar{x}_j) = 0$ for all j, then $q(x) = \sum\limits_{j=1}^{k+1} q(\bar{x}_j) l_j(x) = 0$, which is impossible for a non-constant polynomial $q(x)$. On the other hand, $\sum_{j=1}^{k+1} \bar{\omega}_j q(\bar{x}_j) = \int_0^1 q(x)dx = 0$. Thus we have $m'' < 0 < M''$. So we get

$$q(x) \le \sum_{j=1}^{k+1} |q(\bar{x}_j)||l_j(x)| < \max\{M'', -m''\} \sum_{j=1}^{k+1} |l_j(x)|.$$

Thus $M' < \max\{M'', -m''\} \max\limits_{x\in[0,1]} \sum\limits_{j=1}^{k+1} |l_j(x)| = \max\{M'', -m''\}\Lambda_{k+1}[0,1]$. So we have

$$m' \le m'' < 0 < M' < \max\{M'', -m''\}\Lambda_{k+1}[0,1].$$

Without loss of generality, assume $q(\bar{x}_1) = \max_j q(\bar{x}_j) = M''$. Since $\sum\limits_{j=1}^{k+1} \bar{\omega}_j q(\bar{x}_j) =$

0, we get $\bar{\omega}_1 M'' = \bar{\omega}_1 q(\bar{x}_1) = -\sum\limits_{j=2}^{k+1} \bar{\omega}_j q(\bar{x}_j) \leq -\sum\limits_{j=2}^{k+1} \bar{\omega}_j m'' = -m'' \sum\limits_{j=2}^{k+1} \bar{\omega}_j$, thus

$$\frac{M''}{-m''} \leq \frac{1}{\bar{\omega}_1} \sum_{j=2}^{k+1} \bar{\omega}_j \leq \frac{1}{\min\limits_j \bar{\omega}_j} \sum_{j=2}^{k+1} \bar{\omega}_j \leq \frac{1 - \min\limits_j \bar{\omega}_j}{\min\limits_j \bar{\omega}_j}.$$

Therefore,

$$0 < \frac{M'}{-m'} \leq \frac{\max\{M'', -m''\}\Lambda_{k+1}[0,1]}{-m''} \leq \max\left\{\frac{M''}{-m''}, 1\right\} \Lambda_{k+1}[0,1] \leq \frac{1 - \min\limits_j \bar{\omega}_j}{\min\limits_j \bar{\omega}_j} \Lambda_{k+1}[0,1],$$

where $\frac{1-\min\limits_j \bar{\omega}_j}{\min\limits_j \bar{\omega}_j} = \frac{1-\frac{1}{(k+1)k}}{\frac{1}{(k+1)k}} = k^2 + k - 1$.

Remark 2. We can replace Gauss-Lobatto quadrature rule by other $(k+1)$-point quadrature rules with positive weights in Lemma 4. The proof in Lemma 4 is constructive with the constant C_k given explicitly. But to extend the proof in Lemma 4 to a generic cell in higher dimensions, we need to construct a proper quadrature rule with positive weights. On the other hand, the proof Lemma 3 can be easily extended to any kind of cells in higher dimensions.

In practice, the limiter (15) is not very interesting since evaluating maximum and minimum of high order polynomials in (15b) is computationally demanding especially in high dimensions. Fortunately, we only need to enforce bounds at some points in (13) or (14). A practical limiter is (15) with M_j and m_j redefined as

$$\widetilde{p}_j(x) = \theta\left[p_j(x) - \overline{p}_j\right] + \overline{p}_j, \quad \theta = \min\left\{1, \left|\frac{M - \overline{p}_j}{M_j - \overline{p}_j}\right|, \left|\frac{m - \overline{p}_j}{m_j - \overline{p}_j}\right|\right\}, \tag{16a}$$

$$M_j = \max_{x \in S_j} p_j(x), m_j = \min_{x \in S_j} p_j(x), \tag{16b}$$

where S_j are Gauss-Lobatto quadrature points (9). The simplified limiter (16) was first used in Zhang and Shu (2010b) to enforce the sufficient conditions (13) in high order finite volume and DG schemes with monotone fluxes solving scalar conservation laws.

To enforce (14), we can first compute $p_j(x_j^*) = \frac{1}{1-2\widehat{\omega}}(\overline{u}_j^n - \widehat{\omega}_1 u_{j-\frac{1}{2}}^+ - \widehat{\omega}_N u_{j+\frac{1}{2}}^-)$ then use a more relaxed limiter with M_j and m_j redefined as

$$\widetilde{p}_j(x) = \theta\left[p_j(x) - \overline{p}_j\right] + \overline{p}_j, \quad \theta = \min\left\{1, \left|\frac{M - \overline{p}_j}{M_j - \overline{p}_j}\right|, \left|\frac{m - \overline{p}_j}{m_j - \overline{p}_j}\right|\right\}, \tag{17a}$$

$$M_j = \max_{\widehat{x}_j^1, \widehat{x}_j^N, \widehat{x}_j^*} p_j(x), m_j = \min_{\widehat{x}_j^1, \widehat{x}_j^N, \widehat{x}_j^*} p_j(x). \tag{17b}$$

Since (15) is a more stringent limiter than (16) and (17), Theorem 1 also applies to the simplified limiters (16) and (17).

We remark that it is straightforward to define an optimal limiter in terms of accuracy as an optimization problem, i.e., finding a polynomial $\widetilde{p}_j(x)$ to minimize $\|\widetilde{p}_j(x) - p_j(x)\|$ under the constraints $\int_{I_j} \widetilde{p}_j(x)\,dx = \int_{I_j} p_j(x)\,dx$ and $\widetilde{p}_j(\widehat{x}_j^\mu) \in [m, M]$. But solving these optimization problems accurately is much more computationally demanding. See Guba et al. (2014) for such a limiter.

2.4. SSP high order time discretizations

For high order time discretizations, we can use a strong stability preserving (SSP) Runge-Kutta or multistep method Gottlieb et al. (2009), which is a convex combination of several formal forward Euler steps. For example, let $\frac{d}{dt}u_h = \mathcal{L}(u_h)$ denote a semi-discrete scheme with high order spatial discretizations by a finite volume or DG method, then the third order SSP Runge-Kutta method is given by,

$$\begin{aligned}
u_h^{(1)} &= u_h^n + \Delta t\mathcal{L}(u_h^n),\\
u_h^{(2)} &= \tfrac{3}{4}u_h^n + \tfrac{1}{4}(u_h^{(1)} + \Delta t\mathcal{L}(u_h^{(1)})),\\
u_h^{n+1} &= \tfrac{1}{3}u_h^n + \tfrac{2}{3}(u_h^{(2)} + \Delta t\mathcal{L}(u_h^{(2)})).
\end{aligned} \tag{18}$$

If the forward Euler (8) is bound-preserving, then so are the high order SSP methods due to the convex combinations. Early works using SSP methods to construct high order bound-preserving schemes include Perthame and Shu (1996).

To render a high order finite volume or DG scheme bound-preserving, we should use a SSP time discretizations such as (18) and a monotone flux $\widehat{f}$. Then in each time stage in a Runge-Kutta method or each time step of a multi-step method, we should use the simple limiter (16) or (17).

2.5. Extensions and applications

All discussions in this section can be extended to high dimensions in a straightforward way, see Zhang and Shu (2010b); Zhang et al. (2012); Zhang and Shu (2011a). This framework has been widely used to construct high order schemes which are positivity-preserving for important problems including compressible Euler equations Zhang and Shu (2010c, 2012b, 2011b); Wang et al. (2012), passive convection such as 2D incompressible Euler equations Zhang and Shu (2010b); Zhang et al. (2012), shallow water equations Xing et al. (2010); Xing and Shu (2011); Xing and Zhang (2013), MHD equations Cheng et al. (2013b), Vlasov-Boltzmann equations Cheng et al. (2012), Vlasov-Poisson system Heath et al. (2012); Qiu and Shu (2011b); de Dios et al. (2012); Cheng et al. (2013a), Lagrangian and semi-Lagrangian schemes Rossmanith and Seal (2011); Qiu and Shu (2011a); Guo et al. (2014); Cheng and Shu (2014); Vilar et al. (2016a,b), curvilinear coordinates Endeve et al. (2015), population models Zhang et al. (2011), transport on a sphere Zhang and Nair (2012), extended magnetohydrodynamics equations Zhao et al. (2014), Fokker-Planck equations Liu and Yu (2014), spray and particle transport Larat et al. (2012), pressureless Euler system Yang et al. (2013), relativistic hydrodynamics Qin et al. (2016), streamer discharge simulation Zhuang and Zeng (2014), turbulent cosmology flows Zhu et al. (2013), and other interesting applications Du et al. (2015); Franquet and Perrier (2012); Sabat et al. (2014), etc. Entropy bound-preserving schemes are constructed in Zhang and Shu (2012a); Lv and Ihme (2015). Realizability-preserving DG and WENO schemes are constructed in Olbrant et al. (2012); Alldredge and Schneider (2015); Schneider et al. (2016) for moment closures of kinetic equations. See Zhang (2016) for recent progress on preserving positivity of density and pressure (or internal energy) in compressible Navier-Stokes equations.

3. Bound-preserving flux limiters

In this section, we briefly review the Flux-corrected type bound-preserving limiters within the framework of high order conservative approximation of the scalar conservation laws in one dimension (5).

3.1. Basic idea and framework

The bound-preserving flux limiting approach is to seek a convex combination of the first order monotone flux with the high order flux, in the hope of that such combination can achieve both bound-preserving property high order accuracy under certain conditions, e.g. some mild time step constraint. To achieve this goal, the numerical fluxes have to be modified subject to both bound and accuracy constraint. The foundation of this family of approach can be found in Boris and Book (1973) and improved by Zalesak (1979). The original Boris-Book-Zalesak method is to enforce (3) thus such schemes are at most second order accurate around smooth extrema. The flux limiting approach by Xu (2014); Hu et al. (2013); Xiong et al. (2016) can be viewed as a generalization of the flux-corrected transport method to the high order schemes for scalar conservation laws and compressible Euler systems. For scalar conservation laws, only (2) is enforced in Xu (2014), which makes high order accuracy possible. To illustrate the basic idea, for simplicity, we consider the conservative high order finite difference approximation of (5) with uniform spatial discretization. The main goal here is to achieve the bound-preserving property (2). Fully discretized conservative high order approximation of the 1D scalar conservation law (5) generally assumes the form of

$$u_j^{n+1} = u_j^n - \lambda(\hat{H}_{j+\frac{1}{2}} - \hat{H}_{j-\frac{1}{2}}) \tag{19}$$

with numerical fluxes $\hat{H}_{j\pm\frac{1}{2}}$ obtained from high order spatial reconstruction and temporal integration. For example, the third order SSP Runge-Kutta temporal integration (18) can be written as

$$\begin{aligned}
u_j^{(1)} &= u_j^n + \Delta t \mathcal{L}(u^n), \\
u_j^{(2)} &= u_j^n + \Delta t (\frac{1}{4}\mathcal{L}(u^n) + \frac{1}{4}\mathcal{L}(u^{(1)})), \\
u_j^{n+1} &= u_j^n + \Delta t (\frac{1}{6}\mathcal{L}(u^n) + \frac{2}{3}\mathcal{L}(u^{(2)}) + \frac{1}{6}\mathcal{L}(u^{(1)})).
\end{aligned} \tag{20}$$

We note that the second term on the right hand side of equation (20) approximates $\int_{t^n}^{t^{n+1}} \mathcal{L}(u(\tau))d\tau$. Here $\mathcal{L}(u^n) \doteq -\frac{1}{\Delta x}(\hat{H}^n_{j+\frac{1}{2}} - \hat{H}^n_{j-\frac{1}{2}})$ is the high order spatial derivative approximation and $\hat{H}^n_{j+\frac{1}{2}}$ is the numerical flux obtained from high order reconstruction from u^n Shu and Osher (1988). Similarly, $\hat{H}^{(1)}_{j+\frac{1}{2}}$ and $\hat{H}^{(2)}_{j+\frac{1}{2}}$ are the numerical fluxes reconstructed from the intermediate values $u^{(1)}$ and $u^{(2)}$. It is obvious that the equation (20) is consistent with the description by the form (19) when

$$\hat{H}_{j+\frac{1}{2}} \doteq \frac{1}{6}\hat{H}^n_{j+\frac{1}{2}} + \frac{2}{3}\hat{H}^{(2)}_{j+\frac{1}{2}} + \frac{1}{6}\hat{H}^{(1)}_{j+\frac{1}{2}}.$$

Based on the formulation (19), the parametrized bound-preserving flux limiters are proposed to replace the original high order flux $\hat{H}_{j+\frac{1}{2}}$ by the modified one

$$\tilde{H}_{j+\frac{1}{2}} = \theta_{j+\frac{1}{2}}(\hat{H}_{j+\frac{1}{2}} - \hat{h}_{j+\frac{1}{2}}) + \hat{h}_{j+\frac{1}{2}} \tag{21}$$

so that the bound-preserving property (or positive bounds for the matter of compressible Euler computation) is satisfied by the new scheme

$$u_j^{n+1} = u_j^n - \lambda(\tilde{H}_{j+\frac{1}{2}} - \tilde{H}_{j-\frac{1}{2}}). \tag{22}$$

Here $\hat{h}_{j+\frac{1}{2}}$ is the previously discussed first order monotone flux which preserves the global upper and lower bound. To complete the modification of the fluxes, it requires identifying the parameters $\theta_{j+\frac{1}{2}}$. Through (22), the bound-preserving constraint $m \leq u_j^{n+1} \leq M$ induces a group of linear inequalities that $\theta_{j+\frac{1}{2}}$'s have to satisfy. In Section 3.2, we provide the steps for finding such parameters. We refer to the earlier work Zalesak (1979); Xu (2014); Liang and Xu (2014); Xiong et al. (2013) for the detailed process to decouple those inequalities for $\theta_{j+\frac{1}{2}}$'s explicitly in multi-dimensions.

The flux limiting procedure described above is investigated by Xiong et al. (2013) with application to high order finite difference computation of incompressible flow problems. Different from the *successive* bound-preserving limiting procedure proposed by Xu (2014) for internal stages of a third order SSP Runge-Kutta method, this approach is very general in the sense that it can be applied to any high order method with explicit RK or Lax-Wendroff type integration. It is much simpler to implement. Moreover, the time step restriction to ensure both bound-preserving property and high order accuracy in both space and time is less severe compared with that proposed by Xu (2014).

The parametrized flux limiter has been applied to the semi-Lagrange finite difference WENO computation of Vlasov equations Xiong et al. (2014). It is also generalized to high order finite difference WENO schemes solving compressible Euler equations to preserve positivity of quantities such as density, pressure and energy by Xiong et al. (2016) and solving MHD equations within constraint transport framework Christlieb et al. (2015b). It is also implemented on unstructured mesh Christlieb et al. (2015a). Compared with the polynomial scaling limiters Zhang and Shu (2010b); Zhang et al. (2013), the advantage of the flux limiting approach is for high order finite difference solving conservation laws and for high order schemes solving convection diffusion problems as evidenced by Jiang and Xu (2013); Wu and Tang (2015); Jiang et al. (2015); Xiong et al. (2015). The convenience for solving the convection-diffusion equation is due to that its high order approximation can be cast into the conservative form (21).

However, the major difficulty is to prove that the original high order accuracy is not comprised by the flux limiters in general. Error analysis is performed Xiong et al. (2013) to prove the retainment of third order spatial and temporal accuracy when the high order flux is limited toward a first order local Lax-Friedrich flux or Godunov flux. The proof relies on tedious algebraic calculation and verification. Proof is not available for more general cases such as high order spatial and temporal accuracy for general high order schemes, high dimensional conservation laws, and for the incompressible flow. In the absence of systematic analysis, extensive numerical evidences are provided Xiong et al. (2013) to show the retainment of high order accuracy with chosen CFL numbers.

3.2. Decoupling for the flux limiting parameters

Since identifying the limiting parameters is the main component of the family of the Flux-corrected type numerical methods, we include the detailed procedure of designing $\theta_{j+\frac{1}{2}}$ here. The description follows Xu (2014) for the convenience of discussing the flux limiters for the high order methods with the global bound. For each $\theta_{j+\frac{1}{2}}$ limiting the numerical flux $\hat{H}_{j+\frac{1}{2}}$, we are looking for upper bounds $\Lambda_{-\frac{1}{2},I_j}$ and $\Lambda_{+\frac{1}{2},I_j}$ from the need of keeping u_j^{n+1} within $[m, M]$. Consequently,

$$\theta_{j+\frac{1}{2}} \in [0, \Lambda_{+\frac{1}{2},I_j}] \cap [0, \Lambda_{-\frac{1}{2},I_{j+1}}], \quad \forall j \tag{23}$$

provides a sufficient condition for the scheme to preserve the bound. Let

$$\Gamma_j^M = M - u_j + \lambda(\hat{h}_{j+\frac{1}{2}} - \hat{h}_{j-\frac{1}{2}}), \quad \Gamma_j^m = m - u_j + \lambda(\hat{h}_{j+\frac{1}{2}} - \hat{h}_{j-\frac{1}{2}}),$$

then from the bound-preserving property of a first order monotone scheme,

$$\Gamma_j^M \geq 0, \quad \Gamma_j^m \leq 0.$$

To ensure $u_j^{n+1} \in [m, M]$ with $\tilde{H}_{j+\frac{1}{2}}$ as in equation (22), it is equivalent to require

$$\begin{aligned} \lambda\theta_{j-\frac{1}{2}}(\hat{H}_{j-\frac{1}{2}} - \hat{h}_{j-\frac{1}{2}}) - \lambda\theta_{j+\frac{1}{2}}(\hat{H}_{j+\frac{1}{2}} - \hat{h}_{j+\frac{1}{2}}) - \Gamma_j^M &\leq 0, \qquad (24)\\ \lambda\theta_{j-\frac{1}{2}}(\hat{H}_{j-\frac{1}{2}} - \hat{h}_{j-\frac{1}{2}}) - \lambda\theta_{j+\frac{1}{2}}(\hat{H}_{j+\frac{1}{2}} - \hat{h}_{j+\frac{1}{2}}) - \Gamma_j^m &\geq 0. \qquad (25) \end{aligned}$$

The discussion is case by case based on the sign of

$$F_{j\pm\frac{1}{2}} \doteq \hat{H}_{j\pm\frac{1}{2}} - \hat{h}_{j\pm\frac{1}{2}}.$$

1. Assume

$$\theta_{j-\frac{1}{2}} \in [0, \Lambda^M_{-\frac{1}{2},I_j}], \quad \theta_{j+\frac{1}{2}} \in [0, \Lambda^M_{+\frac{1}{2},I_j}],$$

   where $\Lambda^M_{-\frac{1}{2},I_j}$ and $\Lambda^M_{+\frac{1}{2},I_j}$ are designed to preserve the upper bound by equation (24).

(a) If $F_{j-\frac{1}{2}} \leq 0$ and $F_{j+\frac{1}{2}} \geq 0$,

$$(\Lambda^M_{-\frac{1}{2},I_j}, \Lambda^M_{+\frac{1}{2},I_j}) = (1, 1).$$

(b) If $F_{j-\frac{1}{2}} \leq 0$ and $F_{j+\frac{1}{2}} < 0$,

$$(\Lambda^M_{-\frac{1}{2},I_j}, \Lambda^M_{+\frac{1}{2},I_j}) = (1, \min(1, \frac{\Gamma^M_j}{-\lambda F_{j+\frac{1}{2}} + \epsilon})).$$

(c) If $F_{j-\frac{1}{2}} > 0$ and $F_{j+\frac{1}{2}} \geq 0$,

$$(\Lambda^M_{-\frac{1}{2},I_j}, \Lambda^M_{+\frac{1}{2},I_j}) = (\min(1, \frac{\Gamma^M_j}{\lambda F_{j-\frac{1}{2}} + \epsilon}), 1).$$

(d) If $F_{j-\frac{1}{2}} > 0$ and $F_{j+\frac{1}{2}} < 0$,

- If equation (24) is satisfied with $(\theta_{j-\frac{1}{2}}, \theta_{j+\frac{1}{2}}) = (1, 1)$, then

$$(\Lambda^M_{-\frac{1}{2},I_j}, \Lambda^M_{+\frac{1}{2},I_j}) = (1, 1).$$

- If equation (24) is not satisfied with $(\theta_{j-\frac{1}{2}}, \theta_{j+\frac{1}{2}}) = (1, 1)$, then

$$(\Lambda^M_{-\frac{1}{2},I_j}, \Lambda^M_{+\frac{1}{2},I_j}) = (\frac{\Gamma^M_j}{\lambda F_{j-\frac{1}{2}} - \lambda F_{j+\frac{1}{2}} + \epsilon}, \frac{\Gamma^M_j}{\lambda F_{j-\frac{1}{2}} - \lambda F_{j+\frac{1}{2}} + \epsilon}).$$

2. Similarly assume

$$\theta_{j-\frac{1}{2}} \in [0, \Lambda^m_{-\frac{1}{2},I_j}], \quad \theta_{j+\frac{1}{2}} \in [0, \Lambda^m_{+\frac{1}{2},I_j}],$$

where $\Lambda^m_{-\frac{1}{2},I_j}$ and $\Lambda^m_{+\frac{1}{2},I_j}$ are designed to preserve the lower bound by equation (25).

(a) If $F_{j-\frac{1}{2}} \geq 0$ and $F_{j+\frac{1}{2}} \leq 0$,

$$(\Lambda^m_{-\frac{1}{2},I_j}, \Lambda^m_{+\frac{1}{2},I_j}) = (1, 1).$$

(b) If $F_{j-\frac{1}{2}} \geq 0$ and $F_{j+\frac{1}{2}} > 0$,

$$(\Lambda^m_{-\frac{1}{2},I_j}, \Lambda^m_{+\frac{1}{2},I_j}) = (1, \min(1, \frac{\Gamma^m_j}{-\lambda F_{j+\frac{1}{2}} - \epsilon})).$$

(c) If $F_{j-\frac{1}{2}} < 0$ and $F_{j+\frac{1}{2}} \leq 0$,

$$(\Lambda^m_{-\frac{1}{2},I_j}, \Lambda^m_{+\frac{1}{2},I_j}) = (\min(1, \frac{\Gamma^m_j}{\lambda F_{j-\frac{1}{2}} - \epsilon}), 1).$$

(d) If $F_{j-\frac{1}{2}} < 0$ and $F_{j+\frac{1}{2}} > 0$,

- If equation (25) is satisfied with $(\theta_{j-\frac{1}{2}}, \theta_{j+\frac{1}{2}}) = (1,1)$, then

$$(\Lambda^m_{-\frac{1}{2},I_j}, \Lambda^m_{+\frac{1}{2},I_j}) = (1,1).$$

- If equation (25) is not satisfied with $(\theta_{j-\frac{1}{2}}, \theta_{j+\frac{1}{2}}) = (1,1)$, then

$$(\Lambda^m_{-\frac{1}{2},I_j}, \Lambda^m_{+\frac{1}{2},I_j}) = (\frac{\Gamma^m_j}{\lambda F_{j-\frac{1}{2}} - \lambda F_{j+\frac{1}{2}} - \epsilon}, \frac{\Gamma^m_j}{\lambda F_{j-\frac{1}{2}} - \lambda F_{j+\frac{1}{2}} - \epsilon}).$$

The parameter ϵ can be chosen slightly greater than machine error to avoid division by 0. Notice that the range of $\theta_{j+\frac{1}{2}}$ (23) is determined by the need to ensure both the upper bound (24) and the lower bound (25) of numerical solutions in both cell I_j and its neighboring cells. Thus the locally defined limiting parameter is given as

$$\theta_{j+\frac{1}{2}} = \min(\Lambda_{+\frac{1}{2},I_j}, \Lambda_{-\frac{1}{2},I_{j+1}}), \tag{26}$$

with $\Lambda_{+\frac{1}{2},I_j} = \min(\Lambda^M_{+\frac{1}{2},I_j}, \Lambda^m_{+\frac{1}{2},I_j}), \quad \Lambda_{-\frac{1}{2},I_{j+1}} = \min(\Lambda^M_{-\frac{1}{2},I_{j+1}}, \Lambda^m_{-\frac{1}{2},I_{j+1}})$. The modified flux in equation (21) with the $\theta_{j+\frac{1}{2}}$ designed above ensures the maximum principle. Such modified flux is consistent since it is a convex combination ($\theta_{j+\frac{1}{2}} \in [0,1]$) of a high order flux $\hat{H}_{j+\frac{1}{2}}$ with the first order flux $\hat{h}_{j+\frac{1}{2}}$. The modified scheme (21) is still in the conservative form.

4. Concluding remarks

We have reviewed two approaches for enforcing bound-preserving property in high order schemes. While the approach in Section 2 is simple and effective, it works most well for finite volume and DG schemes solving conservation laws. For finite difference schemes, this approach can also apply to, e.g. compressible Euler equations to maintain positivity Zhang and Shu (2012b), but it has restrictions in order to keep the original high order accuracy. For convection-diffusion equations, this approach works for general DG methods to second order accuracy on unstructured triangular meshes Zhang et al. (2013), and to third order accuracy for a special class of DG methods (the direct DG method) Liu and Yu (2014); Yan (2015); Chen et al. (2016).

The second approach in Section 3 can also be applied to finite volume and DG schemes for conservation laws, but their real advantage is for finite difference schemes solving conservation laws and schemes for solving convection-diffusion equations, for which the first approach has rather severe restrictions as mentioned above.

For spectral methods, see Liu et al. (2016) for a globally defined sweeping limiter to enforce bound-preserving property.

Acknowledgments

The authors are indebted to Prof. Chi-Wang Shu for invaluable discussions, inspirations and encouragement. Xiangxiong Zhang is grateful to Prof. Mark Ainsworth for inspirations on the proof of Lemma 4. Xiangxiong Zhang is supported by the NSF grant DMS-1522593, and Zhengfu Xu is supported by the NSF grant DMS-1316662.

References

Alldredge, G., Schneider, F., 2015. A realizability-preserving discontinuous Galerkin scheme for entropy-based moment closures for linear kinetic equations in one space dimension. Journal of Computational Physics 295, 665–684.

Batten, P., Lambert, C., Causon, D., 1996. Positively Conservative High-resolution Convection Schemes for Unstructured Elements. International Journal for Numerical Methods in Engineering 39 (11), 1821–1838.

Bell, J. B., Dawson, C. N., Shubin, G. R., 1988. An unsplit, higher order Godunov method for scalar conservation laws in multiple dimensions. Journal of Computational Physics 74 (1), 1–24.

Boris, J. P., Book, D. L., 1973. Flux-corrected transport. I. SHASTA, A fluid transport algorithm that works. Journal of computational physics 11 (1), 38–69.

Bouchut, F., 2004. An antidiffusive entropy scheme for monotone scalar conservation laws. Journal of Scientific Computing 21 (1), 1–30.

Cai, X., Zhang, X., Qiu, J., 2015. Positivity-Preserving High Order Finite Volume HWENO Schemes for Compressible Euler Equations. Journal of Scientific Computing, 1–20.
URL http://dx.doi.org/10.1007/s10915-015-0147-8

Chavent, G., Cockburn, B., 1989. The local projection P0-P1 discontinuous-Galerkin finite element method for scalar conservation laws. RAIRO-Modélisation mathématique et analyse numérique 23 (4), 565–592.

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

Cheng, J., Shu, C.-W., 2014. Positivity-preserving Lagrangian scheme for multi-material compressible flow. Journal of Computational Physics 257, 143–168.

Cheng, Y., Gamba, I., Proft, J., 2012. Positivity-preserving discontinuous Galerkin schemes for linear Vlasov-Boltzmann transport equations. Mathematics of Computation 81 (277), 153–190.

Cheng, Y., Gamba, I. M., Morrison, P. J., 2013a. Study of conservation and recurrence of Runge–Kutta discontinuous Galerkin schemes for Vlasov–Poisson systems. Journal of Scientific Computing 56 (2), 319–349.

Cheng, Y., Li, F., Qiu, J., Xu, L., 2013b. Positivity-preserving dg and central DG methods for ideal MHD equations. Journal of Computational Physics 238, 255–280.

Christlieb, A. J., Liu, Y., Tang, Q., Xu, Z., 2015a. High order parametrized maximum-principle-preserving and positivity-preserving weno schemes on unstructured meshes. Journal of Computational Physics 281, 334–351.

Christlieb, A. J., Liu, Y., Tang, Q., Xu, Z., 2015b. Positivity-preserving finite difference weighted eno schemes with constrained transport for ideal magnetohydrodynamic equations. SIAM Journal on Scientific Computing 37 (4), A1825–A1845.

Cockburn, B., Lin, S.-Y., Shu, C.-W., 1989. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: one-dimensional systems. Journal of Computational Physics 84 (1), 90–113.

Cockburn, B., Shu, C.-W., 1989. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. II. general framework. Mathematics of Computation 52 (186), 411–435.

Colella, P., 1990. Multidimensional upwind methods for hyperbolic conservation laws. Journal of Computational Physics 87 (1), 171–200.

de Dios, B. A., Carrillo, J. A., Shu, C.-W., 2012. Discontinuous Galerkin methods for the multi-dimensional Vlasov–Poisson problem. Mathematical Models and Methods in Applied Sciences 22 (12), 1250042.

Du, J., Shu, C.-W., Zhang, M., 2015. A simple weighted essentially non-oscillatory limiter for the correction procedure via reconstruction (CPR) framework. Applied Numerical Mathematics 95, 173–198.

Endeve, E., Hauck, C. D., Xing, Y., Mezzacappa, A., 2015. Bound-Preserving Discontinuous Galerkin Methods for Conservative Phase Space Advection in Curvilinear Coordinates. Journal of Computational Physics 287, 151–183.

Franquet, E., Perrier, V., 2012. Runge–Kutta discontinuous Galerkin method for interface flows with a maximum preserving limiter. Computers & Fluids 65, 2–7.

Gottlieb, S., Ketcheson, D., Shu, C.-W., 2009. High order strong stability preserving time discretizations. Journal of Scientific Computing 38 (3), 251–289.

Guba, O., Taylor, M., St-Cyr, A., 2014. Optimization-based limiters for the spectral element method. Journal of Computational Physics 267, 176–195.

Guermond, J.-L., Nazarov, M., 2014. A maximum-principle preserving C0 finite element method for scalar conservation equations. Computer Methods in Applied Mechanics and Engineering 272, 198–213.

Guermond, J.-L., Nazarov, M., Popov, B., Yang, Y., 2014. A second-order maximum principle preserving Lagrange finite element technique for nonlinear scalar conservation equations. SIAM Journal on Numerical Analysis 52 (4), 2163–2182.

Guo, W., Nair, R. D., Qiu, J.-M., 2014. A Conservative Semi-Lagrangian Discontinuous Galerkin Scheme on the Cubed Sphere. Monthly Weather Review 142 (1), 457–475.

Harten, A., 1983. High resolution schemes for hyperbolic conservation laws. Journal of computational physics 49 (3), 357–393.

Harten, A., Engquist, B., Osher, S., Chakravarthy, S. R., 1987. Uniformly high order accurate essentially non-oscillatory schemes, III. In: Upwind and High-Resolution Schemes. Springer, pp. 218–290.

Heath, R., Gamba, I. M., Morrison, P. J., Michler, C., 2012. A discontinuous Galerkin method for the Vlasov–Poisson system. Journal of Computational Physics 231 (4), 1140–1174.

Hu, X. Y., Adams, N. A., Shu, C.-W., 2013. Positivity-preserving method for high-order conservative schemes solving compressible Euler equations. Journal of Computational Physics 242, 169–180.

Hubbard, M., 1999. Multidimensional slope limiters for MUSCL-type finite volume schemes on unstructured grids. Journal of Computational Physics 155 (1), 54–74.

Jiang, G.-S., Shu, C.-W., 1995. Efficient implementation of weighted ENO schemes. Tech. rep., DTIC Document.

Jiang, G.-S., Tadmor, E., 1998. Nonoscillatory central schemes for multidimensional hyperbolic conservation laws. SIAM Journal on Scientific Computing 19 (6), 1892–1917.

Jiang, Y., Shu, C.-W., Zhang, M., 2015. High-order finite difference weno schemes with positivity-preserving limiter for correlated random walk with density-dependent turning rates. Mathematical Models and Methods in Applied Sciences 25 (08), 1553–1588.

Jiang, Y., Xu, Z., 2013. Parametrized maximum principle preserving limiter for finite difference weno schemes solving convection-dominated diffusion equations. SIAM Journal on Scientific Computing 35 (6), A2524–A2553.

Kurganov, A., Tadmor, E., 2000. New high-resolution central schemes for nonlinear conservation laws and convection–diffusion equations. Journal of Computational Physics 160 (1), 241–282.

Larat, A., Massot, M., Vié, A., et al., 2012. A stable, robust and high order accurate numerical method for Eulerian simulation of spray and particle transport on unstructured meshes. Annual Research Briefs, Center for Turbulence Research, Stanford University, 205–216.

Liang, C., Xu, Z., 2014. Parametrized maximum principle preserving flux limiters for high order schemes solving multi-dimensional scalar hyperbolic conservation laws. Journal of Scientific Computing 58 (1), 41–60.

Liu, H., Yu, H., 2014. Maximum-Principle-Satisfying Third Order Discontinuous Galerkin Schemes for Fokker–Planck Equations. SIAM Journal on Scientific Computing 36 (5), A2296–A2325.

Liu, X.-D., 1993. A maximum principle satisfying modification of triangle based adapative stencils for the solution of scalar hyperbolic conservation laws. SIAM journal on numerical analysis 30 (3), 701–716.

Liu, X.-D., Osher, S., 1996. Nonoscillatory high order accurate self-similar maximum principle satisfying shock capturing schemes I. SIAM Journal on Numerical Analysis 33 (2), 760–779.

Liu, X.-D., Osher, S., Chan, T., 1994. Weighted essentially non-oscillatory schemes. Journal of computational physics 115 (1), 200–212.

Liu, Y., Cheng, Y., Shu, C.-W., 2016. A simple bound-preserving sweeping technique for conservative numerical approximations, submitted to Journal of Scientific Computing.

Lv, Y., Ihme, M., 2015. Entropy-bounded discontinuous Galerkin scheme for Euler equations. Journal of Computational Physics 295, 715–739.

Olbrant, E., Hauck, C. D., Frank, M., 2012. A realizability-preserving discontinuous Galerkin method for the M1 model of radiative transfer. Journal of Computational Physics 231 (17), 5612–5639.

Osher, S., Chakravarthy, S., 1984. High resolution schemes and the entropy condition. SIAM Journal on Numerical Analysis 21 (5), 955–984.

Perthame, B., Shu, C.-W., 1996. On positivity preserving finite volume schemes for Euler equations. Numerische Mathematik 73 (1), 119–130.

Qin, T., Shu, C.-W., Yang, Y., 2016. Bound-preserving discontinuous galerkin methods for relativistic hydrodynamics. Journal of Computational Physics 315, 323–347.

Qiu, J.-M., Shu, C.-W., 2011a. Conservative high order semi-Lagrangian finite difference WENO methods for advection in incompressible flow. Journal of Computational Physics 230 (4), 863–889.

Qiu, J.-M., Shu, C.-W., 2011b. Positivity preserving semi-Lagrangian discontinuous Galerkin formulation: theoretical analysis and application to the Vlasov–Poisson system. Journal of Computational Physics 230 (23), 8386–8409.

Rossmanith, J. A., Seal, D. C., 2011. A positivity-preserving high-order semi-Lagrangian discontinuous Galerkin scheme for the Vlasov–Poisson equations. Journal of Computational Physics 230 (16), 6203–6232.

Sabat, M., Larat, A., Vié, A., Massot, M., 2014. On the development of high order realizable schemes for the Eulerian simulation of disperse phase flows: a convex-state preserving Discontinuous Galerkin method. The Journal of Computational Multiphase Flows 6 (3), 247–270.

Sanders, R., 1988. A third-order accurate variation nonexpansive difference scheme for single nonlinear conservation laws. Math. Comp 51, 535–558.

Schneider, F., Alldredge, G., Kall, J., 2016. A realizability-preserving high-order kinetic scheme using WENO reconstruction for entropy-based moment closures of linear kinetic equations in slab geometry. Kinetic and Related Models 6, 193–215.

Shu, C.-W., 2012. A Brief Survey on High Order Accurate Maximum Principle Satisfying and Positivity Preserving Discontinuous Galerkin and Finite Volume Schemes for Conservation Laws. In: Fifth International Congress of Chinese Mathematicians. Vol. 51. American Mathematical Soc., p. 747.

Shu, C.-W., Osher, S., 1988. Efficient implementation of essentially non-oscillatory shock-capturing schemes. Journal of Computational Physics 77 (2), 439–471.

Vilar, F., Shu, C.-W., Maire, P.-H., 2016a. Positivity-preserving cell-centered Lagrangian schemes for multi-material compressible flows: From first-order to high-orders. Part I: The one-dimensional case. Journal of Computational Physics 312, 385–415.

Vilar, F., Shu, C.-W., Maire, P.-H., 2016b. Positivity-preserving cell-centered Lagrangian schemes for multi-material compressible flows: From first-order to high-orders. Part II: The two-dimensional case. Journal of Computational Physics 312, 416 – 442.

Wang, C., Zhang, X., Shu, C.-W., Ning, J., 2012. Robust high order discontinuous Galerkin schemes for two-dimensional gaseous detonations. Journal of Computational Physics 231, 653–665.

Wu, K., Tang, H., 2015. High-order accurate physical-constraints-preserving finite difference weno schemes for special relativistic hydrodynamics. Journal of Computational Physics 298, 539–564.

Xing, Y., Shu, C.-W., 2011. High-order finite volume WENO schemes for the shallow water equations with dry states. Advances in Water Resources 34 (8), 1026–1038.

Xing, Y., Zhang, X., 2013. Positivity-preserving well-balanced discontinuous Galerkin methods for the shallow water equations on unstructured triangular meshes. Journal of Scientific Computing 57 (1), 19–41.

Xing, Y., Zhang, X., Shu, C.-W., 2010. Positivity-preserving high order well-balanced discontinuous Galerkin methods for the shallow water equations. Advances in Water Resources 33 (12), 1476–1493.

Xiong, T., Qiu, J.-M., Xu, Z., 2013. A parametrized maximum principle preserving flux limiter for finite difference RK-WENO schemes with applications in incompressible flows . Journal of Computational Physics 252, 310–331.

Xiong, T., Qiu, J.-M., Xu, Z., 2015. High order maximum-principle-preserving discontinuous galerkin method for convection-diffusion equations. SIAM Journal on Scientific Computing 37 (2), A583–A608.

Xiong, T., Qiu, J.-M., Xu, Z., 2016. Parametrized positivity preserving flux limiters for the high order finite difference weno scheme solving compressible euler equations. Journal of Scientific Computing 67 (3), 1066–1088.

Xiong, T., Qiu, J.-M., Xu, Z., Christlieb, A., 2014. High order maximum principle preserving semi-lagrangian finite difference weno schemes for the vlasov equation. Journal of Computational Physics 273, 618–639.

Xu, Z., 2014. Parametrized maximum principle preserving flux limiters for high order schemes solving hyperbolic conservation laws: one-dimensional scalar problem. Mathematics of Computation 83 (289), 2213–2238.

Yan, J., 2015. Maximum Principle Satisfying Direct discontinuous Galerkin method and its vairation for convection diffusion equations, submitted for publication.

Yang, Y., Wei, D., Shu, C.-W., 2013. Discontinuous Galerkin method for Krauses consensus models and pressureless Euler equations. Journal of Computational Physics 252, 109–127.

Zalesak, S. T., 1979. Fully multidimensional flux-corrected transport algorithms for fluids. Journal of computational physics 31 (3), 335–362.

Zhang, R., Zhang, M., Shu, C.-W., 2011. High order positivity-preserving finite volume WENO schemes for a hierarchical size-structured population model. Journal of Computational and Applied Mathematics 236 (5), 937–949.

Zhang, X., 2016. On positivity-preserving high order discontinuous Galerkin schemes for compressible Navier-Stokes equations, submitted to Journal of Computational Physics.

Zhang, X., Shu, C.-W., 2010a. A genuinely high order total variation diminishing scheme for one-dimensional scalar conservation laws. SIAM Journal on Numerical Analysis 48 (2), 772–795.

Zhang, X., Shu, C.-W., 2010b. On maximum-principle-satisfying high order schemes for scalar conservation laws. Journal of Computational Physics 229 (9), 3091–3120.

Zhang, X., Shu, C.-W., 2010c. On positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes. Journal of Computational Physics 229 (23), 8918–8934.

Zhang, X., Shu, C.-W., 2011a. Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science 467 (2134), 2752–2776.

Zhang, X., Shu, C.-W., 2011b. Positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations with source terms. Journal of Computational Physics 230 (4), 1238–1248.

Zhang, X., Shu, C.-W., 2012a. A minimum entropy principle of high order schemes for gas dynamics equations. Numerische Mathematik 121 (3), 545–563.

Zhang, X., Shu, C.-W., 2012b. Positivity-preserving high order finite difference weno schemes for compressible euler equations. Journal of Computational Physics 231 (5), 2245–2258.

Zhang, X., Xia, Y., Shu, C.-W., 2012. Maximum-principle-satisfying and positivity-preserving high order discontinuous Galerkin schemes for conservation laws on triangular meshes. Journal of Scientific Computing 50 (1), 29–62.

Zhang, Y., Nair, R. D., 2012. A nonoscillatory discontinuous Galerkin transport scheme on the cubed sphere. Monthly Weather Review 140 (9), 3106–3126.

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

Zhao, X., Yang, Y., Seyler, C. E., 2014. A positivity-preserving semi-implicit discontinuous Galerkin scheme for solving extended magnetohydrodynamics equations. Journal of Computational Physics 278, 400–415.

Zhu, W., Feng, L.-l., Xia, Y., Shu, C.-W., Gu, Q., Fang, L.-Z., 2013. Turbulence in the Intergalactic Medium: Solenoidal and Dilatational Motions and the Impact of Numerical Viscosity. The Astrophysical Journal 777 (1), 48.

Zhuang, C., Zeng, R., 2014. A positivity-preserving scheme for the simulation of streamer discharges in non-attaching and attaching gases. Communications in Computational Physics 15 (01), 153–178.