Monotonicity and Bound-Preserving in High Order Accurate Methods for Convection Diffusion Equations

Xiangxiong Zhang

Math Dept, Purdue Univeristy

USTC online lecture, August 2020

Monotonicity in low order discretization

The second order finite difference $[D_{xx}u]_i = \frac{1}{\Delta x^2}(u_{i-1} - 2u_i + u_{i+1})$ is monotone:

- Forward Euler for $u_t = u_{xx}$:

$$u^{n+1} = u^n + \Delta t D_{xx} u^n \Longrightarrow u_i^{n+1} = u_{i-1}^n + (1 - 2\frac{\Delta t}{\Delta x^2})u_i^n + u_{i+1}^n$$

  Monotonicity means that u^{n+1} is a convex combination of u^n if $\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2}$.

- Backward Euler for $u_t = u_{xx}$:

$$u^{n+1} = u^n + \Delta t D_{xx} u^{n+1} \Longrightarrow u^{n+1} = (I - \Delta t D_{xx})^{-1} u^n$$

  A matrix A is called monotone if its inverse has non-negative entries ($A^{-1} \geq 0$). For second order FD, we have $(I - \Delta t D_{xx})^{-1} \geq 0$.

For convection $u_t + u_x = 0$, the upwind scheme is monotone if $\frac{\Delta t}{\Delta x} \leq 1$:

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x}(u_i^n - u_{i-1}^n) = (1 - \frac{\Delta t}{\Delta x})u_i^n + \frac{\Delta t}{\Delta x} u_{i-1}^n$$

Monotonicity implies the Discrete Maximum Principle (DMP):

$$\min_i u_i^n \leq u_i^{n+1} \leq \max_i u_i^n.$$

Compressible Navier-Stokes Equations in Gas Dynamics

$$\begin{pmatrix} \rho \\ \rho\mathbf{u} \\ E \end{pmatrix}_t + \nabla\cdot\begin{pmatrix} \rho\mathbf{u} \\ \rho\mathbf{u}\otimes\mathbf{u} + p\mathbb{I} \\ (E+p)\mathbf{u} \end{pmatrix} = \nabla\cdot\begin{pmatrix} 0 \\ \boldsymbol{\tau} \\ \mathbf{u}\boldsymbol{\tau} - \mathbf{q} \end{pmatrix},$$

$$E = \frac{1}{2}\rho\|\mathbf{u}\|^2 + \rho e.$$

Equation of the State: $p = (\gamma - 1)\rho e,$

Newtonian approximation $\boldsymbol{\tau} = \eta(\boldsymbol{\nabla}\mathbf{u} + \boldsymbol{\nabla}^t\mathbf{u}) + (\eta_b - \frac{2}{3}\eta)(\nabla\cdot\mathbf{u})\mathbb{I},$

Fourier's Law $\mathbf{q} = -\kappa\nabla T.$

Sutherland formula $\eta = \dfrac{C_1\sqrt{T}}{1 + C_2/T}.$

Stokes hypothesis $\eta_b = 0.$

Two-dimensional dimensionless compressible NS equations

$$\begin{pmatrix}\rho\\ \rho u\\ \rho v\\ E\end{pmatrix}_t+\begin{pmatrix}\rho u\\ \rho u^2+p\\ \rho uv\\ (E+p)u\end{pmatrix}_x+\begin{pmatrix}\rho v\\ \rho vu\\ \rho v^2+p\\ (E+p)v\end{pmatrix}_y=\frac{1}{\mathrm{Re}}\begin{pmatrix}0\\ \tau_{xx}\\ \tau_{yx}\\ \tau_{xx}u+\tau_{yx}v+\frac{\gamma e_x}{\mathrm{Pr}}\end{pmatrix}_x+\frac{1}{\mathrm{Re}}\begin{pmatrix}0\\ \tau_{xy}\\ \tau_{yy}\\ \tau_{xy}u+\tau_{yy}v+\frac{\gamma e_y}{\mathrm{Pr}}\end{pmatrix}_y,$$

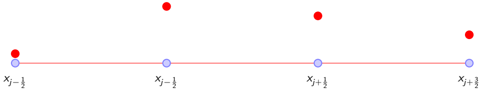
$$e=\frac{1}{\rho}\left(E-\frac{1}{2}\rho u^2-\frac{1}{2}\rho v^2\right),\quad p=(\gamma-1)\rho e,$$

$$\begin{aligned}\tau_{xx}&=\frac{4}{3}u_x-\frac{2}{3}v_y,\\ \tau_{xy}&=\tau_{yx}=u_y+v_x,\\ \tau_{yy}&=\frac{4}{3}v_y-\frac{2}{3}u_x.\end{aligned}$$

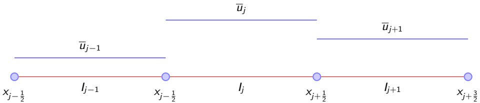
Popular Spatial Discretizations in Semi-Discrete Methods for Time-Dependent Problems

Methods of approximating/representing a smooth function in a finite dimensional space:

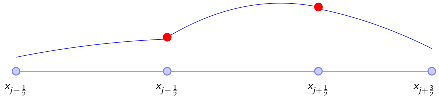
- Spectral Method: $u(x) = \sum_{i=1}^{N} a_i \phi_i(x)$ where $\phi_i(x)$ form a basis of L^2 functions, e.g., trigonometric functions or polynomials.
- Finite Difference:



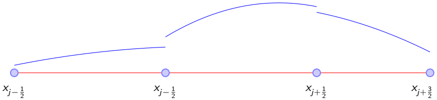
- Finite Volume:



- Continuous Galerkin:



- Discontinuous Galerkin:



High order accurate methods: beyond piecewise linear or second order accuracy.

Stability: Compressible Euler Equations in Gas Dynamics

$$\begin{pmatrix} \rho \\ m \\ E \end{pmatrix}_t + \begin{pmatrix} m \\ \rho u^2 + p \\ (E+p)u \end{pmatrix}_x = 0,$$

with

$$m = \rho u, \quad E = \frac{1}{2}\rho u^2 + \rho e, \quad p = (\gamma - 1)\rho e.$$

The speed of sound is given by $c = \sqrt{\gamma p/\rho}$ and the three eigenvalues of the Jacobian are $u, u \pm c$.

If either $\rho < 0$ or $p < 0$, then the sound speed is imaginary and the system is no longer hyperbolic. Thus the initial value problem is ill-posed. This is why it is computationally unstable.

Mach 10 shock passing a triangle

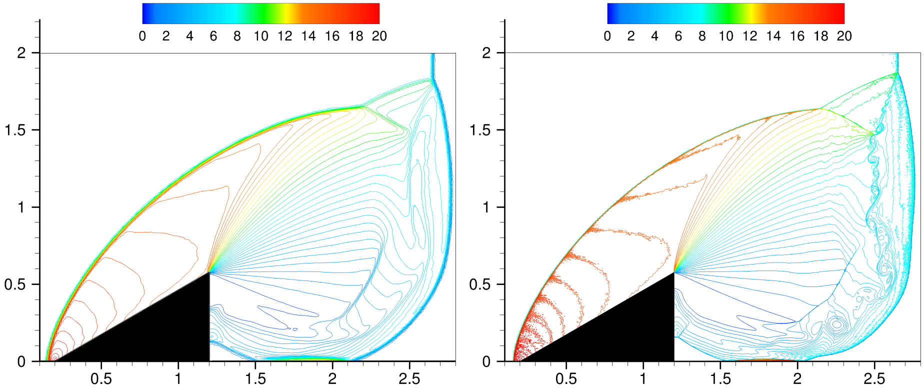
Plot of Density. Numerical result of our positivity-preserving **sixth** order accurate Runge-Kutta Discontinuous Galerkin scheme on unstructured triangular meshes. Navier-Stokes, Re=1000.

1. The higher the shock speed is, the lower the density/pressure will be after the diffraction.
2. Third order Runge-Kutta DG scheme with TVB limiter is not stable due to loss of positivity.
3. *Ad Hoc* tricks to preserve positivity in a high order code:

- Replace negative ρ or p by positive ones. (loss of conservation; blows up at a later time)
- Use a first order positivity preserving scheme in trouble cells.

Positivity-Preserving RKDG, Re=1000

Positivity-preserving explicit high order DG for compressible Navier-Stokes in X.Z. 2017:



(a) Locally replace high order solutions by P1 in trouble cells.

(b) P7, mesh size $\frac{1}{160}$.

High Speed Flow in Astrophysical Jets Modelling: Mach 2000

Plot of Density. Scales are logarithmic. Numerical result of our positivity-preserving third order DG.
The second order MUSCL, high order ENO/WENO and DG schemes are unstable for this example.

Objective

- Positivity itself is easy to achieve.
- The challenge is how to achieve positivity without losing certain constraints:
  1. Conservation: it's hard to enforce internal energy positivity without losing total energy conservation for Navier-Stokes even for a second order accurate scheme.
  2. Accuracy: many conservative efficient method loses high order accuracy.
  3. Efficiency and practical concern: a global optimization type limiter is usually unacceptable; cost effective, multi-dimensions, unstructured meshes, parallelizability and etc.
- The key is to exploit monotonicity (up to some sense) in high order schemes. Advantage of using monotonicity: easier extension to more general and demanding applications.

Upwind scheme for solving $u_t + u_x = 0$:

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x}(u_i^n - u_{i-1}^n) = \left(1 - \frac{\Delta t}{\Delta x}\right) u_i^n + \frac{\Delta t}{\Delta x} u_{i-1}^n.$$

Plan

- What and why: explore monotonicity toward positivity-preserving.

- Part I: Weak Monotonicity in High Order Schemes with Explicit Time Discretizations:
  1. X.Z. and Shu, 2010: bound-preserving for scalar equations.
  2. X.Z. and Shu, 2010: positivity-preserving in compressible Euler equations.
  3. X.Z., 2017: positivity-preserving in compressible Navier-Stokes equations.

  Finite Volume and Discontinuous Galerkin Schemes on unstructured meshes.

$$\text{Navier-Stokes} \longrightarrow \text{Euler} \longrightarrow \text{Scalar Convection} \longrightarrow u_t + u_x = 0.$$

- Part II: Weak Monotonicity in some Finite Difference schemes.
- Part III: Monotonicity in finite difference implementation of finite element method with backward euler for solving $u_t = \nabla \cdot (a(\mathbf{x})\nabla u)$:
  - Hao Li and X.Z., 2020: finite difference implementation of continuous finite element method with quadratic basis on rectangular meshes.

Bound Preserving for Scalar Conservation Laws

Consider the initial value problem

$$u_t + \nabla \cdot \mathbf{F}(u) = 0, u(\mathbf{x}, 0) = u_0(\mathbf{x}), \mathbf{x} \in \mathbb{R}^n$$

for which the unique entropy solution $u(x, t)$ satisfies

$$\min_{\mathbf{x}} u(\mathbf{x}, t_0) \leq u(\mathbf{x}, t) \leq \max_{\mathbf{x}} u(\mathbf{x}, t_0), \quad \forall t \geq t_0. \qquad \text{Maximum Principle}$$

In particular,

$$\min_{\mathbf{x}} u_0(\mathbf{x}) = m \leq u(\mathbf{x}, t) \leq M = \max_{\mathbf{x}} u_0(\mathbf{x}). \qquad \text{Bound Preserving}$$

It is also a desired property for numerical solutions due to

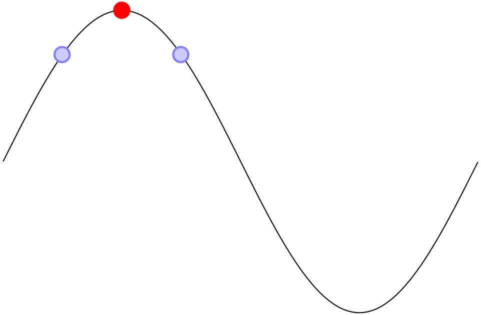
1. Physical meaning: vehicle density (traffic flow), mass percentage (pollutant transport), probability distribution (Boltzmann equation) and etc.
2. Stability for systems: positivity of density and pressure (gas dynamics), water height (shallow water equations), particle density for describing electrical discharges (a convection-dominated system) and etc.

For numerical schemes, this is a completely DIFFERENT problem from discrete maximum principle in solving elliptic equations.

Scalar Equations

- IVP: $u_t + f(u)_x = 0, u(x,0) = u_0(x)$.
- Maximum Principle (Bound Preserving): $u(x,t) \in [m, M]$ where $m = \min u_0(x), M = \min u_0(x)$.
- For finite difference, any scheme satisfying $\min_j u_j^n \leq u_j^{n+1} \leq \max_j u_j^n$ can be at most first order accurate.

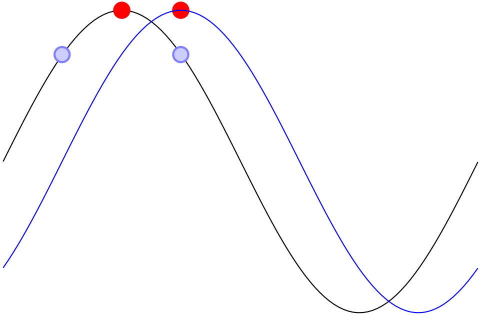
Harten's Counter Example: consider $u_t + u_x = 0, \quad u(x,0) = \sin x$. Put the grids in a way such that $x = \frac{\pi}{2}$ is in the middle of two grid points.



Scalar Equations

- IVP: $u_t + f(u)_x = 0, u(x,0) = u_0(x)$.
- Maximum Principle (Bound Preserving): $u(x,t) \in [m, M]$ where $m = \min u_0(x), M = \min u_0(x)$.
- For finite difference, any scheme satisfying $\min_j u_j^n \leq u_j^{n+1} \leq \max_j u_j^n$ can be at most first order accurate.

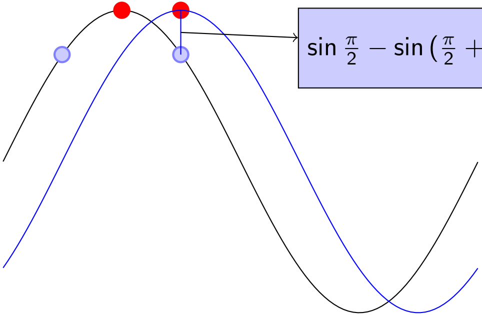
Harten's Counter Example: consider $u_t + u_x = 0, \quad u(x,0) = \sin x$. Put the grids in a way such that $x = \frac{\pi}{2}$ is in the middle of two grid points.



Scalar Equations

- IVP: $u_t + f(u)_x = 0, u(x,0) = u_0(x)$.
- Maximum Principle (Bound Preserving): $u(x,t) \in [m, M]$ where $m = \min u_0(x), M = \min u_0(x)$.
- For finite difference, any scheme satisfying $\min_j u_j^n \le u_j^{n+1} \le \max_j u_j^n$ can be at most first order accurate.

Harten's Counter Example: consider $u_t + u_x = 0, \quad u(x,0) = \sin x$. Put the grids in a way such that $x = \frac{\pi}{2}$ is in the middle of two grid points.



$$\sin\frac{\pi}{2} - \sin\left(\frac{\pi}{2} + \frac{\Delta x}{2}\right) = \frac{1}{8}\Delta x^2 + O(\Delta x^3)$$

Bound-Preserving Schemes

1. First order monotone schemes.
2. FD/FV schemes satisfying $\min_j u_j^n \leq u_j^{n+1} \leq \max_j u_j^n$: can have any formal order of accuracy in the monotone region but are only first order accurate around the extrema. E.g.,
   - Conventional total-variation-diminishing (TVD) schemes.
   - High Resolution schemes such as the MUSCL scheme.
3. FV schemes satisfying $\min_x u^n(x) \leq u^{n+1}(x) \leq \max_x u^n(x)$:
   - R. Sanders, 1988: a third order finite volume scheme for 1D.
   - X.Z. and Shu, 2009: higher order (up to 6th) extension of Sanders scheme.
   - Liu and Osher, 1996: a third order FV scheme for 1D (can be proven bound-preserving only for linear equations).
   - Noelle, 1998; Kurganov and Petrova, 2001: 2D generalization of *Liu and Osher*.
   - All schemes in this category use the exact time evolution.
4. Just bound-preserving: $m \leq u^{n+1} \leq M$. Practical/popular high order schemes are NOT bound-preserving. It was unknown previously how to construct a high order bound-preserving scheme for 2D nonlinear equations.

Explicit Time Discretization: SSP Runge-Kutta or Multi-Step Method

High order strong stability preserving (SSP) Runge-Kutta or multi-step method is a convex combination of several forward Euler schemes. E.g., the third order SSP Runge-Kutta method for solving $u_t = F(u)$ is given by

$$\begin{aligned} u^{(1)} &= u^n + \Delta t F(u^n) \\ u^{(2)} &= \frac{3}{4}u^n + \frac{1}{4}(u^{(1)} + \Delta t F(u^{(1)}) \\ u^{n+1} &= \frac{1}{3}u^n + \frac{2}{3}(u^{(2)} + \Delta t F(u^{(2)})) \end{aligned}$$

- If the forward Euler is bound-preserving, then so is the high order Runge-Kutta/Multi-Step.
- SSP time discretization has been often used to construct positivity preserving schemes but previous methods are not high order accurate because the high order spatial accuracy are destroyed (or DIFFICULT to justify).

Conservative Eulerian Schemes

Integrate $u_t + f(u)_x = 0$ on an interval $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$, we have

$$\int_{I_j} u_t dx + f\left(u\left(x_{j+\frac{1}{2}}, t\right)\right) - f\left(u\left(x_{j-\frac{1}{2}}, t\right)\right) = 0. \tag{1}$$

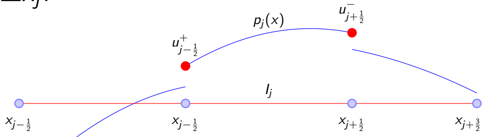
Let $\overline{u}$ denote the cell average. With forward Euler time discretization:

$$\overline{u}_j^{n+1} = \overline{u}_j^n - \frac{\Delta t}{\Delta x_j}\left[f\left(u^n\left(x_{j+\frac{1}{2}}\right)\right) - f\left(u^n\left(x_{j-\frac{1}{2}}\right)\right)\right]. \tag{2}$$

Conservative Schemes: the approximation to the flux $f\left(u^n\left(x_{j+\frac{1}{2}}\right)\right)$ is single-valued even though the approximation to $u^n\left(x_{j-\frac{1}{2}}\right)$ are usually double-valued. The scheme must have the following form,

$$\overline{u}_j^{n+1} = \overline{u}_j^n - \frac{\Delta t}{\Delta x_j}\left[\hat{f}_{j+\frac{1}{2}} - \hat{f}_{j-\frac{1}{2}}\right]. \tag{3}$$

Global Conservation: $\sum_j \overline{u}_j^{n+1}\Delta x_j = \sum_j \overline{u}_j^n \Delta x_j.$



Positivity and Conservation Imply L1-Stability

We insist on using conservative schemes:

1. Lax-Wendroff Theorem: if converging (as mesh sizes go to zero), the converged solution of a conservative scheme is a weak solution.
2. The shock location will be wrong if the conservation is violated.
3. If a scheme is conservative and positivity preserving, then we have L1-stability:

$$\sum_j |\overline{u}_j^{n+1}|\Delta x_j = \sum_j \overline{u}_j^{n+1}\Delta x_j = \sum_j \overline{u}_j^{n}\Delta x_j = \sum_j |\overline{u}_j^{n}|\Delta x_j.$$

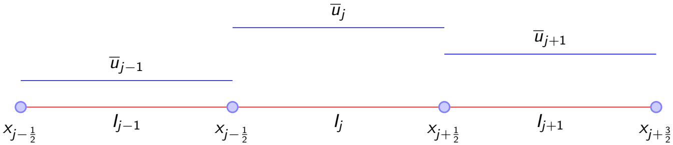
- In Euler equations, if density and pressure are positive, then we have L1-stability for density and total energy.
- Crude replacement of negative values by positive ones is simply unacceptable and unstable because it destroys the local conservation.

First Order Finite Volume Schemes for Compressible Euler

Quite a few first order accurate schemes are positivity preserving for compressible Euler equations:

$$\overline{u}_j^{n+1} = \overline{u}_j^n - \frac{\Delta t}{\Delta x}\left[\hat{f}(\overline{u}_{j-1}^n, \overline{u}_j^n) - \hat{f}(\overline{u}_j^n, \overline{u}_{j+1}^n)\right]$$

- Godunov's Scheme: $\hat{f}$ is the exact solution to the Riemann problem.
- Lax-Friedrichs Scheme: $\hat{f}(u,v) = \frac{1}{2}\left[f(u) + f(v) - \alpha(v-u)\right]$ where $\alpha = \max|f'(u)|$.
- HLLE Scheme: an approximate Riemann solver. Positivity was proved in B. Einfeldt, C.D. Munz, P.L. Roe and B. Sjögreen, JCP, 1991



First Order Schemes for Scalar Conservation Laws

Let $\lambda = \frac{\Delta t}{\Delta x}$, a monotone scheme for $u_t + f(u)_x = 0$ is given by

$$\begin{aligned} u_j^{n+1} &= u_j^n - \lambda \left[\widehat{f}(u_j^n, u_{j+1}^n) - \widehat{f}(u_{j-1}^n, u_j^n) \right] \\ &= H(u_{j-1}^n, u_j^n, u_{j+1}^n). \end{aligned}$$

where the numerical flux $\widehat{f}(\uparrow, \downarrow)$ is monotonically increasing w.r.t. the first variable and decreasing w.r.t. the second variable. E.g., the Lax-Friedrichs flux

$$\widehat{f}(u, v) = \frac{1}{2}(f(u) + f(v) - \alpha(v - u)), \alpha = \max_u |f'(u)|.$$

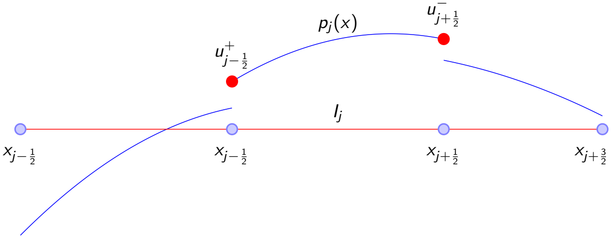
If $m \leq u_j^n \leq M$ for all j, then $H(\uparrow, \uparrow, \uparrow)$ implies

$$m = H(m, m, m) \leq u_j^{n+1} \leq H(M, M, M) = M.$$

High Order Spatial Discretization

$$\overline{u}_j^{n+1} = \overline{u}_j^n - \frac{\Delta t}{\Delta x}\left[\widehat{f}(u^-_{j+\frac{1}{2}}, u^+_{j+\frac{1}{2}}) - \widehat{f}(u^-_{j-\frac{1}{2}}, u^+_{j-\frac{1}{2}})\right].$$

- Finite Volume (FV): given cell averages $\overline{u}_j$ for all j, reconstruct a polynomial $p_j(x)$ on each interval I_j. Evolve only the cell averages in time. Example: high order ENO and WENO schemes.
- Discontinuous Galerkin (DG): find a piecewise polynomial approximation satisfying the integral equation. Evolve all the polynomial $p_j(x)$ in time.



No Straightforward Monotonicity for High Order DG and FV Schemes

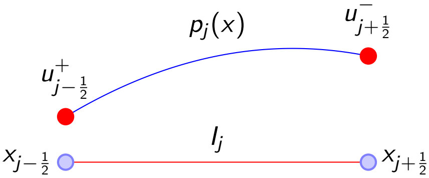
Consider the first order forward Euler time discretization:

$$\begin{aligned}
\overline{u}_j^{n+1} &= \overline{u}_j^n - \lambda \left[\widehat{f}(u^-_{j+\frac{1}{2}}, u^+_{j+\frac{1}{2}}) - \widehat{f}(u^-_{j-\frac{1}{2}}, u^+_{j-\frac{1}{2}}) \right] \\
&= H\left(\overline{u}_j^n, u^-_{j+\frac{1}{2}}, u^+_{j+\frac{1}{2}}, u^-_{j-\frac{1}{2}}, u^+_{j-\frac{1}{2}} \right)
\end{aligned}$$

No Straightforward Monotonicity for High Order DG and FV Schemes

Consider the first order forward Euler time discretization:

$$
\begin{aligned}
\overline{u}_j^{n+1} &= \overline{u}_j^n - \lambda\left[\widehat{f}(u^-_{j+\frac{1}{2}}, u^+_{j+\frac{1}{2}}) - \widehat{f}(u^-_{j-\frac{1}{2}}, u^+_{j-\frac{1}{2}})\right] \\
&= H\left(\overline{u}_j^n, u^-_{j+\frac{1}{2}}, u^+_{j+\frac{1}{2}}, u^-_{j-\frac{1}{2}}, u^+_{j-\frac{1}{2}}\right) \\
&= H(\uparrow, \downarrow, \uparrow, \uparrow, \downarrow)
\end{aligned}
$$

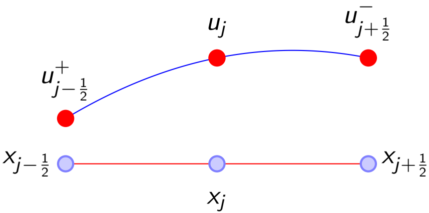


This means $\overline{u}_j^{n+1}$ could be negative even if $\overline{u}_j^n, u^-_{j+\frac{1}{2}}, u^+_{j+\frac{1}{2}}, u^-_{j-\frac{1}{2}}, u^+_{j-\frac{1}{2}}$ are all positive no matter how small the time step is.

Weak Monotonicity for High Order DG and FV Schemes

Consider solving $u_t + u_x = 0$ and upwind flux $\hat{f}(u^-, u^+) = u^-$ and third order accurate schemes:

$$\begin{aligned}
\overline{u}_j^{n+1} &= \overline{u}_j^n - \lambda \left[\widehat{f}(u^-_{j+\frac{1}{2}}, u^+_{j+\frac{1}{2}}) - \widehat{f}(u^-_{j-\frac{1}{2}}, u^+_{j-\frac{1}{2}}) \right] \\
&= \overline{u}_j^n - \lambda u^-_{j+\frac{1}{2}} + \lambda u^-_{j-\frac{1}{2}} \\
&= \left[\frac{1}{6} u^+_{j-\frac{1}{2}} + \frac{2}{3} u_j + \frac{1}{6} u^-_{j+\frac{1}{2}} \right] - \lambda u^-_{j+\frac{1}{2}} + \lambda u^-_{j-\frac{1}{2}} \\
&= H(u^+_{j-\frac{1}{2}}, u_j, u^-_{j+\frac{1}{2}}, u^-_{j-\frac{1}{2}})
\end{aligned}$$



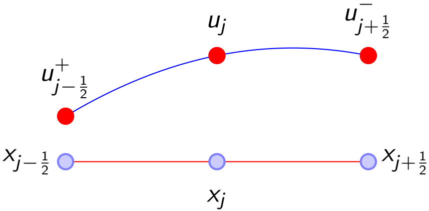
3-point Gauss-Lobatto quadrature is exact for quadratic polynomial $p(x)$:

$$\bar{u} = \frac{1}{\Delta x} \int_{I_j} p(x) dx = \frac{1}{6} u^+_{j-\frac{1}{2}} + \frac{2}{3} u_j + \frac{1}{6} u^-_{j+\frac{1}{2}}.$$

Weak Monotonicity for High Order DG and FV Schemes

Consider solving $u_t + u_x = 0$ and upwind flux $\hat{f}(u^-, u^+) = u^-$ and third order accurate schemes:

$$\begin{aligned}
\overline{u}_j^{n+1} &= \overline{u}_j^n - \lambda\left[\widehat{f}(u^-_{j+\frac{1}{2}}, u^+_{j+\frac{1}{2}}) - \widehat{f}(u^-_{j-\frac{1}{2}}, u^+_{j-\frac{1}{2}})\right] \\
&= \overline{u}_j^n - \lambda u^-_{j+\frac{1}{2}} + \lambda u^-_{j-\frac{1}{2}} \\
&= \left[\frac{1}{6}u^+_{j-\frac{1}{2}} + \frac{2}{3}u_j + \frac{1}{6}u^-_{j+\frac{1}{2}}\right] - \lambda u^-_{j+\frac{1}{2}} + \lambda u^-_{j-\frac{1}{2}} \\
&= H(u^+_{j-\frac{1}{2}}, u_j, u^-_{j+\frac{1}{2}}, u^-_{j-\frac{1}{2}})
\end{aligned}$$



3-point Gauss-Lobatto quadrature is exact for quadratic polynomial $p(x)$:

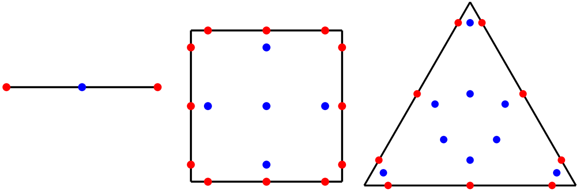
$$\bar{u} = \frac{1}{\Delta x}\int_{I_j} p(x)dx = \frac{1}{6}u^+_{j-\frac{1}{2}} + \frac{2}{3}u_j + \frac{1}{6}u^-_{j+\frac{1}{2}}.$$

Main Result: A Weak Monotonicity for Arbitrarily High Order Schemes

Theorem (X.Z. and Shu, 2010, 2011; X.Z., Xia and Shu 2012)

The sufficient conditions for $\overline{u}^{n+1} \in [m, M]$ are

1. *At time t^n, $p_j(x)$ at points of a special quadrature are in $[m, M]$.*
2. *CFL: $\frac{\Delta t}{\Delta x} \max_u |f'(u)|$ should be less than the smallest weight of this quadrature.*



The special quadrature for quadratic polynomials.

This quadrature is defined by:

1. Quadrature points include all red points.
2. The smallest weight is positive.

- The quadrature is not used for computing.
- $\overline{u}^{n+1}$ is monotone w.r.t. these points.
- We only need the existence of this quadrature and its smallest weight.

Existence of The Special Quadrature

One can construct this quadrature in any dimension:

- 1D: Gauss-Lobatto.
- 2D rectangle: tensor product of Gauss and Gauss-Lobatto.
- 2D triangle (any): Dubinar Transform of the rectangles.
- 2D polygon: union of several triangles.
- 3D tetrahedron.
- Curvilinear element: more quadrature points.

Remarks

1. This quadrature is not used for computing any integral.
2. All we need in computation is the smallest weight, which gives a very natural CFL condition (comparable to the one required by linear stability).

CFL conditions for 1D Discontinuous Galerkin method

Table: The CFL for DG method with polynomial of degree $2 \leq k \leq 5$.

k	The Smallest Weight is $\frac{1}{k(k+1)}$	Linear Stability $\frac{1}{2k+1}$
2	$1/6$	$1/5$
3	$1/6$	$1/7$
4	$1/12$	$1/9$
5	$1/12$	$1/11$

Remarks

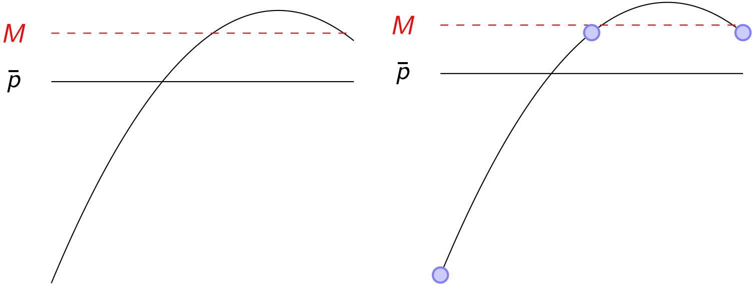
1. The CFL for bound-preserving is sufficient rather than necessary.
2. The CFL needed by bound-preserving is comparable to the one of linear stability.

A Scaling Limiter

Given $p(x)$ with $\bar{p} \in [m, M]$, we need to modify it such that $p(x) \in [m, M]$ for any $x \in I$. Liu and Osher (1996):

$$\widetilde{p}(x) = \theta(p(x) - \bar{p}) + \bar{p}, \qquad \theta = \min\left\{ \left| \frac{M - \bar{p}}{M' - \bar{p}} \right|, \left| \frac{m - \bar{p}}{m' - \bar{p}} \right|, 1 \right\}.$$

where $m' = \min\limits_{x \in I} p(x), M' = \max\limits_{x \in I} p(x)$.

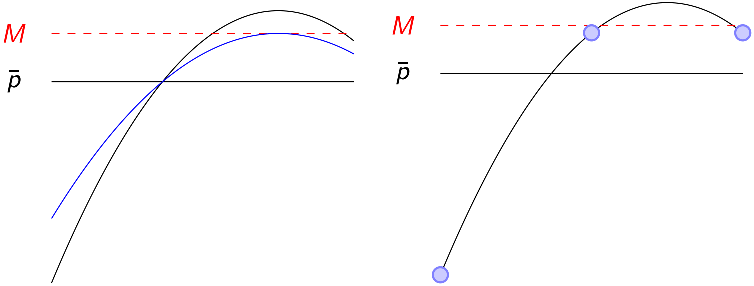


A Scaling Limiter

Given $p(x)$ with $\bar{p} \in [m, M]$, we need to modify it such that $p(x) \in [m, M]$ for any $x \in I$. Liu and Osher (1996):

$$\widetilde{p}(x) = \theta(p(x) - \bar{p}) + \bar{p}, \qquad \theta = \min\left\{ \left| \frac{M - \bar{p}}{M' - \bar{p}} \right|, \left| \frac{m - \bar{p}}{m' - \bar{p}} \right|, 1 \right\}.$$

where $m' = \min\limits_{x \in I} p(x)$, $M' = \max\limits_{x \in I} p(x)$.



A Simple Scaling Limiter

Given $p(x)$, we need to modify it such that $p(x) \in [m, M]$ for any $x \in S$, where S is the set of special quadrature points.

$$\widetilde{p}(x) = \theta(p(x) - \bar{p}) + \bar{p}, \qquad \theta = \min\left\{\left|\frac{M - \bar{p}}{M' - \bar{p}}\right|, \left|\frac{m - \bar{p}}{m' - \bar{p}}\right|, 1\right\}.$$

where $m' = \min_{x \in S} p(x)$, $M' = \max_{x \in S} p(x)$. This limiter is

- Conservative: cell averages are unchanged.
- Cheap to implement: no need to evaluate the extrema.
- High Order Accurate: $p_j(x) - \widetilde{p}_j(x) = \mathcal{O}(\Delta x^{k+1})$ for smooth solutions.

Lemma (X.Z. and Shu, 2010)

$|p(x) - \widetilde{p}(x)| \le C_k |p(x) - u(x)|, \forall x \in I$. The constant C_k depends only on the polynomial degree k and the dimension of the problem.

High Order Bound-Preserving Schemes

A high order bound-preserving scheme can be constructed as follows X.Z. and Shu, 2010:

1. Use high order SSP Runge-Kutta or Multi-Step discretization.
2. Use finite volume or DG spatial discretization with a monotone flux, e.g., Lax-Friedrichs flux.
3. Use the simple limiter in every time stage/step.

- The full scheme is conservative and high order accurate (in the sense of local truncation error).
- Easy to code: add the limiter to a high order FV/DG code.
- Efficiency: we have avoided evaluating the max/min of polynomials. We can also avoid evaluating the redundant blue point values.
- Easy extension to any dimension/mesh.
- The limiter is local thus does not affect the parallelizability at all.

This is the first high order bound-preserving scheme for nonlinear equations in 2D/3D.

Positivity-Preserving: Compressible Euler Equations

$$\begin{pmatrix} \rho \\ m \\ E \end{pmatrix}_t + \begin{pmatrix} m \\ \rho u^2 + p \\ (E+p)u \end{pmatrix}_x = 0,$$

with

$$m = \rho u, \quad E = \frac{1}{2}\rho u^2 + \rho e, \quad p = (\gamma - 1)\rho e.$$

- the set of admissible states is a convex set

$$G = \left\{ \mathbf{w} = \begin{pmatrix} \rho \\ m \\ E \end{pmatrix} \middle| \rho \geq 0, p = (\gamma - 1)\left(E - \frac{1}{2}\frac{m^2}{\rho}\right) \geq 0 \right\}.$$

- If $\rho \geq 0$, the pressure $p(\mathbf{w}) = (\gamma - 1)(E - \frac{1}{2}\frac{m^2}{\rho})$ is a concave function of $\mathbf{w} = (\rho, m, E)$: Jensen's inequality

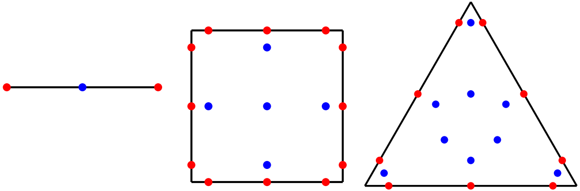
$$p(\lambda_1 \mathbf{w_1} + \lambda_2 \mathbf{w_2}) \geq \lambda_1 p(\mathbf{w_1}) + \lambda_2 p(\mathbf{w_2}).$$

Weak Positivity for Compressible Euler Equations

Theorem (X.Z. and Shu, 2010, 2011)

The sufficient conditions for $\overline{w}^{n+1} \in G$ *are*

1. *At time* t^n, $q_j(x)$ *at points of a special quadrature are in* G.
2. *CFL:* $\frac{\Delta t}{\Delta x} \max(|u| + c)$ *should be less than the smallest weight of this quadrature.*



The special quadrature for quadratic polynomials.

- The weak monotonicity extends to weak positivity for pressure due to Jensen's inequality and positivity-preserving fluxes (Godunov, HLLE, Lax-Friedrichs, kinetic types, etc).
- Similar limiter to enforce the positivity of density and pressure.
- A generic EOS: internal energy is always a concave function.

Why The Weak Monotonicity/Positivity Matters

It answers the following questions:

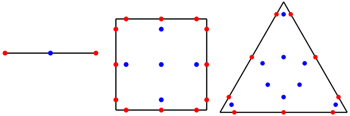
- Is it possible to construct a practical high order conservative positivity-preserving scheme (in what sense, to what extent)?

  YES (rigorous justification for high order local truncation error; arbitrarily high order DG or any Finite Volume scheme)

- For the sake of positivity (robustness), how to properly modify existing high order FV and RKDG codes without destroying conservation or accuracy?

  Add a simple limiter:

  - Easy to code. For each cell, the limiter does not depend on info outside of this cell.
  - Cost of limiter is marginal: we can avoid evaluating redundant blues points.
  - Stringent CFL. But it is not necessary. Enforce it only when negative values emerge.
  - No mesh constraint.



Two-dimensional dimensionless compressible NS equations

$$\begin{pmatrix}\rho\\ \rho u\\ \rho v\\ E\end{pmatrix}_t+\begin{pmatrix}\rho u\\ \rho u^2+p\\ \rho uv\\ (E+p)u\end{pmatrix}_x+\begin{pmatrix}\rho v\\ \rho vu\\ \rho v^2+p\\ (E+p)v\end{pmatrix}_y=\frac{1}{\mathrm{Re}}\begin{pmatrix}0\\ \tau_{xx}\\ \tau_{yx}\\ \tau_{xx}u+\tau_{yx}v+\frac{\gamma e_x}{\mathrm{Pr}}\end{pmatrix}_x+\frac{1}{\mathrm{Re}}\begin{pmatrix}0\\ \tau_{xy}\\ \tau_{yy}\\ \tau_{xy}u+\tau_{yy}v+\frac{\gamma e_y}{\mathrm{Pr}}\end{pmatrix}_y,$$

$$e=\frac{1}{\rho}\left(E-\frac{1}{2}\rho u^2-\frac{1}{2}\rho v^2\right),\quad p=(\gamma-1)\rho e,$$

$$\begin{aligned}\tau_{xx}&=\frac{4}{3}u_x-\frac{2}{3}v_y,\\ \tau_{xy}&=\tau_{yx}=u_y+v_x,\\ \tau_{yy}&=\frac{4}{3}v_y-\frac{2}{3}u_x.\end{aligned}$$

Highly nontrivial to construct second order conservative schemes in 2D/3D preserving the positivity of internal energy without losing conservation of the total energy.

- Grapsas et al., 2015: a second order unconditionally stable scheme

A Toy Problem: $u_t = u_{xx}$

Monotonicity in explicit time stepping:

- Higher order accurate linear schemes are not monotone.

Weak monotonicity of linear schemes in explicit time stepping:

- Weak Monotonicity holds up to second order accuracy in local truncation errors in a FV/DG type scheme, e.g., Y. Zhang, X. Z. and C.-W. Shu 2013, P1 LDG.
- X.Z., Liu and Shu 2012: High order nonconventional FV.
- Chen, Huang and Yan, 2016: third DDG.
- Hao Li and X.Z. 2018: 4th, 6th, 8th order compact finite difference schemes.

Weak Monotonicity for nonlinear discretizations in explicit time stepping:

- Sun, Carrillo and Shu 2017: high order DG for gradient flows.
- Srinivasan, Poggie and X.Z. 2018: high order DG; an additional limiter is needed; constraint on boundary conditions.

Still difficult to generalize it to weak positivity of pressure in NS system.

A Positivity-Preserving Flux

1. We can regard NS system as convection-diffusion

$$\mathbf{U}_t + \nabla \cdot \mathbf{F}^a = \nabla \cdot \mathbf{F}^d.$$

   or formally convection

$$\mathbf{U}_t + \nabla \cdot \mathbf{F} = 0, \quad \mathbf{F} = \mathbf{F}^a - \mathbf{F}^d.$$

2. X.Z. and Shu, JCP 2010: weak positivity holds for high order finite volume scheme

$$\overline{\mathbf{U}}_K^{n+1} = \overline{\mathbf{U}}_K^n - \frac{\Delta t}{|K|} \int_{\partial K} \widehat{\mathbf{F} \cdot \mathbf{n}} \, ds.$$

   if $\widehat{\mathbf{F} \cdot \mathbf{n}}$ is a positivity-preserving flux.
3. X.Z., JCP 2017: a positivity-preserving flux $\widehat{\mathbf{F} \cdot \mathbf{n}}$, which is a nonlinear discretization to the NS diffusion operator.

The Positivity-Preserving Flux in DG Schemes

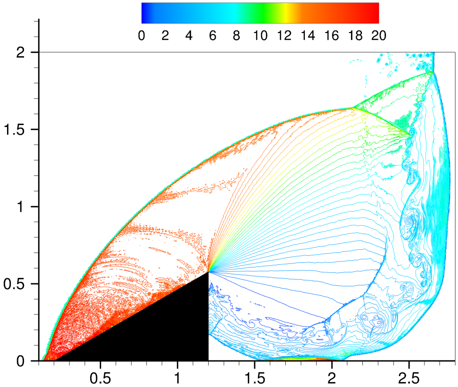
- Quite a few different DG schemes for compressible NS: Bassi and Rebay, 1997; Uranga, Persson, Drela and Peraire, 2009 (Compact DG), Peraire, Nguyen and Cockburn, 2010 (Hybridizable DG), Peraire, Nguyen and Cockburn (Embedded DG), etc.
- DG schemes for $\mathbf{U}_t + \nabla \cdot \mathbf{F}^a = \nabla \cdot \mathbf{F}^d$ take the form

$$\int_K \mathbf{U}_t v \, dV - \int_K \mathbf{F}^a \nabla \cdot v \, ds + \int_{\partial K} v \widehat{\mathbf{F}^a \cdot \mathbf{n}} \, ds = -\int_K \mathbf{F}^d \nabla \cdot v \, ds + \int_{\partial K} v \widehat{\mathbf{F}^d \cdot \mathbf{n}} \, ds$$

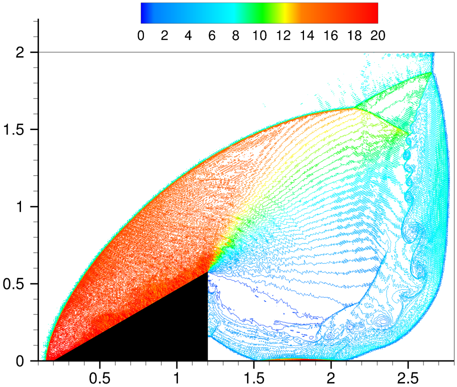
- Bassi and Rebay, 1997: a mixed finite element method, with S approximating ∇U, $\widehat{\mathbf{F}^d \cdot \mathbf{n}}(\mathbf{U}^-, \mathbf{U}^+, \mathbf{S}^-, \mathbf{S}^+) = \frac{1}{2}\left[\mathbf{F}^d(\mathbf{U}^-, \mathbf{S}^-) \cdot \mathbf{n} + \mathbf{F}^d(\mathbf{U}^+, \mathbf{S}^+) \cdot \mathbf{n}\right]$
- The positivity-preserving flux $\widehat{\mathbf{F}^d \cdot \mathbf{n}}(\mathbf{U}^-, \mathbf{U}^+, \mathbf{S}^-, \mathbf{S}^+) = \frac{1}{2}\left[\mathbf{F}^d(\mathbf{U}^-, \mathbf{S}^-) \cdot \mathbf{n} + \mathbf{F}^d(\mathbf{U}^+, \mathbf{S}^+) \cdot \mathbf{n} + \beta(\mathbf{U}^+ - \mathbf{U}^-)\right]$,

$$\beta = \max_{\mathbf{U}^+, \mathbf{U}^-} \frac{1}{2\rho^2 e}\left(\sqrt{\rho^2 |\mathbf{q} \cdot \mathbf{n}|^2 + 2\rho^2 e \|\boldsymbol{\tau} \cdot \mathbf{n}\|_2^2} + \rho |\mathbf{q} \cdot \mathbf{n}|\right)$$

Positivity-Preserving RKDG, Re=∞



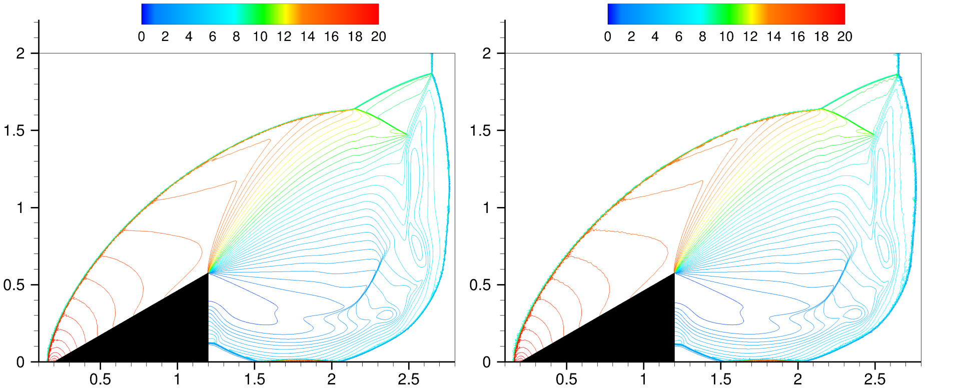
(c) P2, mesh size $\frac{1}{160}$.



(d) P4, mesh size $\frac{1}{80}$.

- Gibbs Phenomenon: higher order schemes are more oscillatory.
- A positivity-preserving scheme can produce highly oscillatory solutions.
- Low artificial viscosity of the positivity-preserving limiter.

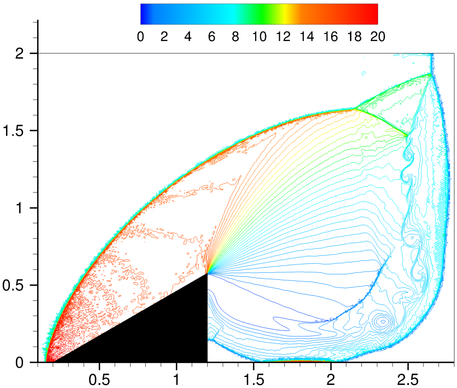
Positivity-Preserving RKDG, Re=100



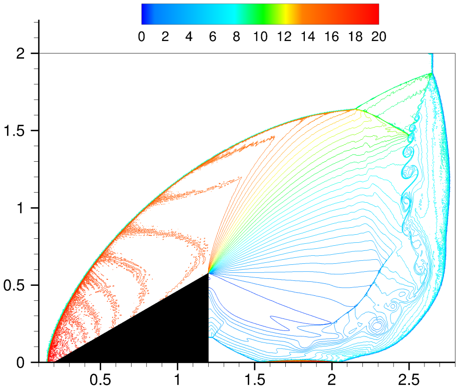
(e) P2, mesh size $\frac{1}{160}$.

(f) P4, mesh size $\frac{1}{80}$.

Positivity-Preserving RKDG, Re=1000

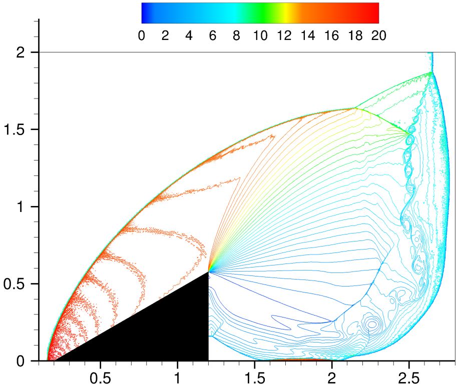


(g) P2, mesh size $\frac{1}{80}$.

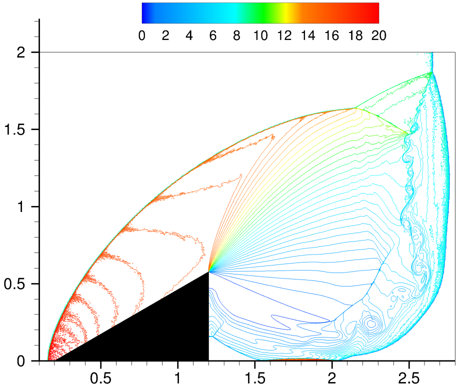


(h) P4, mesh size $\frac{1}{160}$.

Positivity-Preserving RKDG, Re=1000

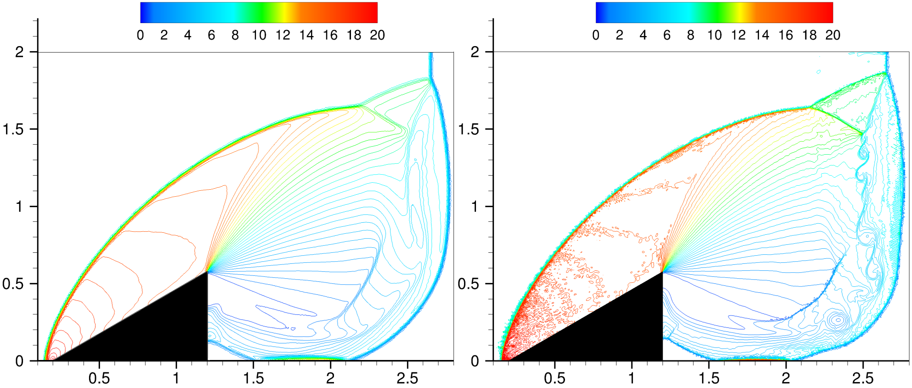


(i) P5, mesh size $\frac{1}{160}$.



(j) P7, mesh size $\frac{1}{160}$.

Low Artificial Dissipation of the positivity-preserving DG Method

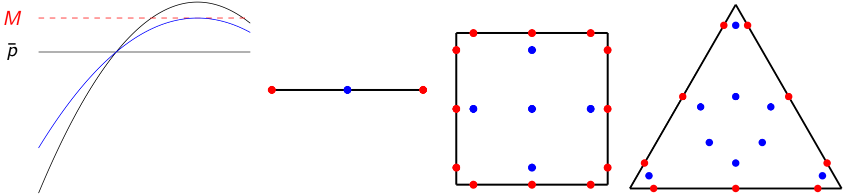


Left: positivity-preserving third order RKDG with TVB limiter (in trouble cells, high order oscillatory polynomials are replaced by linear polynomials). Right: positivity-preserving third order RKDG. Re=1000.

Contributions

This framework to construct positivity-preserving schemes is based on:

- Shu, 1988; Shu and Osher, 1988: Strong Stability Preserving time discretizations.
- Perthame and Shu, 1996: high order FV schemes can be written as a convex combination of several formal first order schemes.
- X.-D. Liu and S. Osher, 1996: the simple scaling limiter.



- X.Z. and Shu, 2010: the weak monotonicity/positivity for $\mathbf{U}_t + \nabla \cdot \mathbf{F} = 0$.
- X.Z., JCP 2017: a positivity-preserving flux for the diffusion operator in compressible NS, which is a nonlinear discretization/approximation to the diffusion operator.

Concluding Remarks of Positivity-Preserving Explicit High Order Schemes for Navier-Stokes

Features of this approach:

- The very first high order schemes for compressible NS that are conservative and positivity-preserving.
- The approach applies to any finite volume schemes: use SSP Runge-Kutta, use positivity-preserving fluxes, then add a simple positivity-preserving limiter.
- Easy extension to 3D, general shapes of computational cells including curved ones.
- It does not affect the parallelizability at all because the positivity-preserving limiter is local to each cell.
- Explicit, CFL $\Delta t = \mathcal{O}(\mathrm{Re}\,\Delta x^2)$; suitable for high Reynolds number.
- It does not depend on EOS, the definition of $\boldsymbol{\tau}$ and $\mathbf{q}$, or how they are approximated. (Stability does not imply convergence.)

Interesting observation: the numerical solutions of high order DG is not oscillatory when the nonlinear diffusion is resolved.

Positivity-Preserving Implicit Schemes for Navier-Stokes

D. Grapsas, R. Herbin, W. Kheriji, J.-C. Latché, 2015:

- MAC type scheme (similar to solving incompressible Navier-Stokes)
- Implicit unconditionally stable scheme
- But only for simplified dimensionless form of the compressible NS system (Laplacian on the internal energy)
- Second order finite difference forms an M-matrix for Laplacian:

$$-D_{xx}u = f, \quad -D_{xx} = \frac{1}{\Delta x^2}\begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -1 & 2 & -1 & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}.$$

A matrix A is called monotone if its inverse has non-negative entries ($A^{-1} \geq 0$). $-D_{xx}$ is an M-matrix thus monotone.

Part II: Weak Monotonicity in Some Finite Difference Schemes

Unfortunately, in general the weak monotonicity does not hold for high order finite difference schemes. However, some finite difference schemes can be perceived as pseudo finite volume schemes thus weak monotonicity holds for an auxiliary variable but not the original variable.

- X. Z. and Shu, 2012: finite difference WENO schemes for compressible Euler equations.
- Hao Li, Xie and X. Z., 2018: compact finite difference for scalar convection diffusion.

Why finite difference: easier implement and lower computational cost thus still preferred on rectangular domains.

The Auxiliary Variable in Finite Difference Schemes

Consider solving $u_t + u_x = 0$. A conservative semi-discrete finite difference scheme can be written as

$$\frac{d}{dt}u_i(t) = -\frac{1}{\Delta x}(\hat{f}_{i+\frac{1}{2}} - \hat{f}_{i-\frac{1}{2}}) \tag{4}$$

where $\frac{1}{\Delta x}(\hat{f}_{i+\frac{1}{2}} - \hat{f}_{i-\frac{1}{2}})$ should be a high order approximation to u_x at x_i.

Assume there is a function $h(x)$ such that $u(x) = \frac{1}{\Delta x}\int_{x-\frac{\Delta x}{2}}^{x+\frac{\Delta x}{2}} h(\xi)d\xi$ then:

- Point values of u are cell averages of $h(x)$:

$$u(x_i) = \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} h(\xi)d\xi = \bar{h}_i$$

- $u_x = \frac{1}{\Delta x}[h(x+\frac{\Delta x}{2}) - h(x-\frac{\Delta x}{2})]$.
- The scheme (4) is a finite volume scheme for $h(x)$:

$$\frac{d}{dt}\bar{h}_i = -\frac{1}{\Delta x}(\hat{f}_{i+\frac{1}{2}} - \hat{f}_{i-\frac{1}{2}})$$

  where $\hat{f}_{i+\frac{1}{2}}$ is a high order approximation to $h(x_{i+\frac{1}{2}})$.

Positivity-Preserving Finite Difference WENO for Compressible Euler

So the finite difference scheme has weak monotonicity for auxiliary variable $h(x)$ (depending on Δx), which is not exactly $u(x)$.

- h will converge to u as Δx goes to zero.
- For a fixed Δx, h has larger maximum and smaller minimum than u.
- If $u \geq 0$, then enforcing positivity for $h(x)$ will destroy high order accuracy.
- If $u > 0$, then , then $h(x) \geq 0$ for small enough Δx thus high order accuracy is possible by preserving positivity of $h(x)$.
- X. Z. and Shu, 2012: positivity is achieved by adding the same simple limiter in Part I for $h(x)$ in finite difference WENO schemes for compressible Euler equations. In gas/fluid dynamics, vacuum state does not make any sense in continuum equations.

Fourth Order Compact Finite Difference

Standard centered finite difference:

$$u_i' = \frac{u_{i+1} - u_{i-1}}{2\Delta x} + \mathcal{O}(\Delta x^2)$$

$$u_i'' = \frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2} + \mathcal{O}(\Delta x^2)$$

Fourth order compact finite difference:

$$\frac{1}{6}u_{i+1}' + \frac{4}{6}u_i' + \frac{1}{6}u_{i-1}' = \frac{u_{i+1} - u_{i-1}}{2\Delta x} + \mathcal{O}(\Delta x^4)$$

$$\frac{1}{12}u_{i+1}'' + \frac{5}{6}u_i'' + \frac{1}{12}u_{i-1}'' = \frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2} + \mathcal{O}(\Delta x^4)$$

A tridiagonal system needs to be solved:

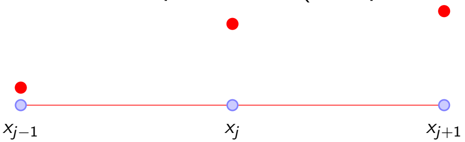
$$\frac{1}{6}\begin{pmatrix} 4 & 1 & & & & 1 \\ 1 & 4 & 1 & & & \\ & 1 & 4 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & 4 & 1 \\ 1 & & & & 1 & 4 \end{pmatrix} \begin{pmatrix} u_1' \\ u_2' \\ u_3' \\ \vdots \\ u_{N-1}' \\ u_N' \end{pmatrix} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{N-1} \\ u_N \end{pmatrix}, \quad W\mathbf{u}' = \mathbf{u}.$$

The Weighting Operator for Convection

If we regard W as an operator mapping a vector to another vector, then

$$(Wu)_j = \frac{1}{6}u_{j+1} + \frac{4}{6}u_j + \frac{1}{6}u_{j-1},$$

which happens to be the Simpson's rule (or 3-point Gauss-Lobatto Rule) in quadrature.



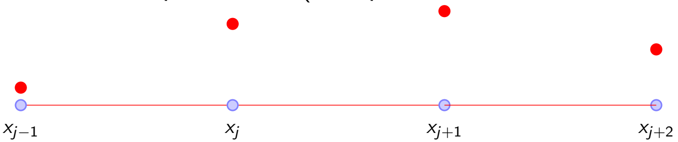
Locally, for each interval $[x_{j-1}x_{j+1}]$, there exists a cubic polynomial $p_j(x)$, obtained through interpolation at $x_{j-1}, x_j, x_{j+1}, x_{j+2}$ (or $x_{j-2}, x_{j-1}, x_j, x_{j+1}$)

The Weighting Operator for Convection

If we regard W as an operator mapping a vector to another vector, then

$$(Wu)_j = \frac{1}{6}u_{j+1} + \frac{4}{6}u_j + \frac{1}{6}u_{j-1},$$

which happens to be the Simpson's rule (or 3-point Gauss-Lobatto Rule) in quadrature.



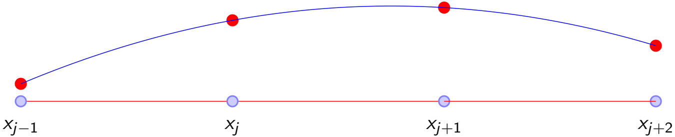
Locally, for each interval $[x_{j-1}x_{j+1}]$, there exists a cubic polynomial $p_j(x)$, obtained through interpolation at $x_{j-1}, x_j, x_{j+1}, x_{j+2}$ (or $x_{j-2}, x_{j-1}, x_j, x_{j+1}$)

The Weighting Operator for Convection

If we regard W as an operator mapping a vector to another vector, then

$$(Wu)_j = \frac{1}{6}u_{j+1} + \frac{4}{6}u_j + \frac{1}{6}u_{j-1},$$

which happens to be the Simpson's rule (or 3-point Gauss-Lobatto Rule) in quadrature.



Locally, for each interval $[x_{j-1}x_{j+1}]$, there exists a cubic polynomial $p_j(x)$, obtained through interpolation at $x_{j-1}, x_j, x_{j+1}, x_{j+2}$ (or $x_{j-2}, x_{j-1}, x_j, x_{j+1}$)

The Fourth Order Compact Finite Difference Scheme for Convection

Let $\overline{u}_i = (Wu)_i = \frac{1}{6}u_{i-1} + \frac{4}{6}u_i + \frac{1}{6}u_{i+1}$. The fourth order compact finite difference for $u_t + f(u)_x = 0$ can be written as

$$\overline{u}_i^{n+1} = \overline{u}_i^n + \frac{\Delta t}{\Delta x}\frac{1}{2}[f(u_{i+1}^n) - f(u_{i-1}^n)],$$

or equivalently

$$u_i^{n+1} = u_i^n + \frac{1}{2}\lambda W^{-1}[f(u_{i+1}^n) - f(u_{i-1}^n)].$$

The weak monotonicity holds under the CFL constraint $\lambda \max_u |f'(u)| \leq \frac{1}{3}$:

$$\begin{aligned}\overline{u}_i^{n+1} &= \frac{1}{6}u_{i-1}^n + \frac{4}{6}u_i^n + \frac{1}{6}u_{i+1}^n + \frac{1}{2}\lambda[f(u_{i+1}^n) - f(u_{i-1}^n)]\\ &= \frac{1}{6}[u_{i-1} - 3\lambda f(u_{i-1}^n)] + \frac{1}{6}[u_{i+1}^n + 3\lambda f(u_{i+1}^n)] + \frac{4}{6}u_i^n\\ &= H(u_{i-1}^n, u_i^n, u_{i+1}^n) = H(\uparrow,\uparrow,\uparrow).\end{aligned}$$

Thus $m \leq u_i^n \leq M$ implies $m = H(m,m,m) \leq \overline{u}_i^{n+1} \leq H(M,M,M) = M$.

Diffusion

$$\frac{1}{12}u''_{i+1} + \frac{5}{6}u''_i + \frac{1}{12}u''_{i-1} = \frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2} + \mathcal{O}(\Delta x^4)$$

Let $\overline{u}_i = (Wu)_i = \frac{1}{12}u_{i-1} + \frac{10}{12}u_i + \frac{1}{12}u_{i+1}$. The fourth order compact finite difference for $u_t = g(u)_{xx}$ can be written as

$$\overline{u}_i^{n+1} = \overline{u}_i^n + \frac{\Delta t}{\Delta x^2}[g(u_{i+1}^n) - 2g(u_i^n) + g(u_{i-1}^n)],$$

Assuming $g'(u) \geq 0$. The weak monotonicity holds under the CFL constraint $\frac{\Delta t}{\Delta x^2} \max_u |f'(u)| \leq \frac{1}{6}$.

Remarks

In general, the weak monotonicity does not hold for high finite volume and DG methods for diffusion, except the third order direct DG method.

Bound-Preserving Compact FD for Scalar Convection Diffusion

- The weak monotonicity can be extended to a convection diffusion equation and 2D/3D.
- Higher Order Accuracy: sixth order and eighth order accurate compact finite difference operators satisfying weak monotonicity for both convection and diffusion can be constructed.
- Given $\bar{u}_i \in [m, M]$, a simple high order limiter can be designed to enforce $u_i \in [m, M]$.
- Inflow-outflow boundary conditions for pure convection: a straightforward fourth order accurate boundary scheme.
- Dirichlet boundary conditions for convection diffusion: a straightforward third order accurate boundary scheme.
- Generalization to Systems?
  Let G be a convex set and u_i denote a vector, then weak positivity $\bar{u}_i \in G$ still holds. But the difficult is on designing the limiter.

Monotonicity for Schemes Solving $u_t + f(u)_x = 0$.

- Godunov Theorem, 1959: a monotonicity preserving scheme is at most first order accurate.
- Harten, Hyman and Lax, 1976: a monotone scheme is at most first order accurate.
- X. Z. and Shu, 2010: arbitrarily high order FV and DG schemes are weakly monotone.
- Hao Li, Xie and X. Z., 2018: 4th, 6th, 8th order compact Finite Difference schemes are weakly monotone.

Monotonicity is not a necessary condition for bound-preserving or positivity-preserving but it is a very convenient tool.

Part III: monotonicity in implicit schemes for diffusion

The second order finite difference $[D_{xx}u]_i = \frac{1}{\Delta x^2}(u_{i-1} - 2u_i + u_{i+1})$ is monotone:

- Forward Euler for $u_t = u_{xx}$:

  $$u^{n+1} = u^n + \Delta t D_{xx} u^n \Longrightarrow u_i^{n+1} = u_{i-1}^n + (1 - 2\frac{\Delta t}{\Delta x^2})u_i^n + u_{i+1}^n$$

  Monotonicity means that u^{n+1} is a convex combination of u^n if $\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2}$.

- Backward Euler for $u_t = u_{xx}$:

  $$u^{n+1} = u^n + \Delta t D_{xx} u^{n+1} \Longrightarrow u^{n+1} = (I - \Delta t D_{xx})^{-1} u^n$$

  A matrix A is called monotone if its inverse has non-negative entries ($A^{-1} \geq 0$). For second order FD, we have $(I - \Delta t D_{xx})^{-1} \geq 0$.

Monotonicity implies the Discrete Maximum Principle (DMP):

$$\min_i u_i^n \leq u_i^{n+1} \leq \max_i u_i^n .$$

Monotonicity in High Order Schemes

- Explicit time discretization: high order SSP Runge-Kutta+ Compact Finite Difference is weakly monotone.
  Hao Li, Xie and X.Z., SINUM 2018.

- Implicit time discretization: backward Euler+FD implementation of Lagrange Q^2 Finite Element Method is monotone for $u_t = \nabla \cdot (a(\mathbf{x})\nabla u)$.
  Hao Li and X.Z. 2020a, Numerische Mathematik.

  1. Xu and Zikatanov 1999: P1 FEM on unstructured meshes is monotone for $-\nabla \cdot (a(\mathbf{x})\nabla u)$.
  2. Höhn and Mittelmann 1981: P2 FEM does not satisfy DMP for $-\Delta u$ on unstructured meshes.
  3. For the Laplacian $-\Delta u$, a few high order schemes are monotone on structured grid:
     - Bramble and Hubbard 1963: 9-point discrete Laplacian.
     - Lorenz 1977: P2 FEM on regular triangular mesh.
  4. No high order schemes had been proven monotonicity for $-\nabla \cdot (a(\mathbf{x})\nabla u)$.

Plan

From now on, we focus on

- $u_t = \nabla \cdot (a \nabla u)$
- $-\nabla \cdot (a \nabla u^{n+1}) + \frac{1}{\Delta t} u^{n+1} = \frac{1}{\Delta t} u^n$
- $-\nabla \cdot (a \nabla u) + cu = f$

1. The finite difference (FD) implementation of Lagrange Q^2 Finite Element Method: a variational difference method.

2. It is fourth order accurate (superconvergence).

3. It is monotone thus satisfies the Discrete Maximum Principle (DMP).

Q^2 FEM for 2D Poisson Equation on a Rectangle

- Consider solving the Poisson equation $-\nabla \cdot (a\nabla u) = f, \quad a(x,y) > 0$ with homogeneous Dirichlet b.c. on a rectangular domain Ω.
- Variational form: seek $u \in H_0^1(\Omega)$ to satisfy
$$A(u,v) = (f,v), \quad \forall v \in H_0^1(\Omega)$$
$$A(u,v) = \iint_\Omega a\nabla u \cdot \nabla v dxdy, \quad (f,v) = \iint_\Omega f v dxdy.$$
- C^0-Q^k finite element: seek $u_h \in V_0^h$ to satisfy
$$A(u_h, v_h) = (f, v_h), \quad \forall v_h \in V_0^h$$
$V_0^h \subset H_0^1(\Omega)$ consists of continuous piecewise Q^k polynomials on a rectangular mesh Ω_h.
- Standard error estimates:
$$\|u - u_h\|_{H^1} = \mathcal{O}(h^k), \quad \|u - u_h\|_{L^2} = \mathcal{O}(h^{k+1})$$

.

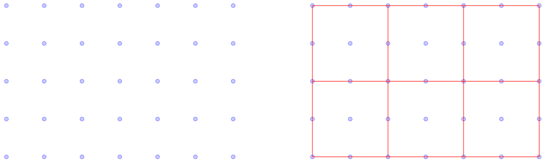
Two Implementations of Q^2 FEM for 2D Elliptic Problems on a Rectangle

1. Replace $a(x, y)$ by its Q^2 interpolant $a_I(x, y)$:

$$\iint_\Omega a_I(x, y)\nabla u_h \cdot \nabla v_h dxdy = (f, v_h).$$

2. C^0-Q^2 FEM with 3×3 GL quadrature is fourth order accurate in the discrete 2-norm over all GL points. This is a FD scheme.

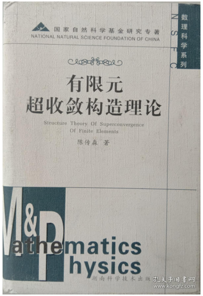
   Ciarlet and Raviart 1972: standard estimates hold if using any quadrature (exact for Q^{2k-1}) for $\iint_\Omega a\nabla u_h \cdot \nabla v_h dxdy$. For Q^2, 3×3 Gauss-Lobatto is enough. It's not even exact for $a \equiv 1$.



- Standard error estimates hold for two implementations.
- They are both fourth order accurate (superconvergence). Superconvergence of function values for 2D variable case: for $k \geq 2$, $u - u_h$ is of order $k + 2$ at all Gauss-Lobatto points over whole domain.

Superconvergence for the original scheme $\iint_\Omega a(x,y)\nabla u_h \cdot \nabla v_h dxdy = (f, v_h)$:

- Superconvergence of function values for 2D variable case: for $k \geq 2$, $u - u_h$ is of order $k+2$ at all Gauss-Lobatto points over whole domain in the discrete 2-norm.
- Original papers:
  - Chen 1980s.
  - Lin, Yan and Zhou 1991.
- Complete rigorous proof can be found in two books in Chinese.



Two Superconvergence Results

For a general elliptic PDE $-\sum\limits_{i=1}^{2} \partial_i \left(\sum\limits_{j=1}^{2} a^{ij} \partial_j u \right) + \sum_{j=1}^{2} b_j \partial_j u + cu = f$:

Hao Li and X.Z. 2020b, JSC: using a third order accurate coefficient a_I in the PDE.

- It looks surprising because of the Q^2 interpolation error $a(x,y) - a_I(x,y) = \mathcal{O}(h^3)$.
- It boils down to the integral of $\iint_\Omega [a(x,y) - a_I(x,y)]$, which is the Gauss Lobatto quadrature error thus one order higher.
- We use standard tools (Bramble-Hilbert Lemma type arguments) thus it can be extended to any Q^k element $k \geq 2$ and 3D.

Hao Li and X.Z. 2020c, JSC: use GL quadrature for integrals (FD implementation).

- It does not look surprising because the quadrature is fourth order accurate.
- Bramble-Hilbert Lemma does not work: it gives sharp quadrature error estimate on each cell but not on whole domain.
- Superconvergence techniques + explicit quadrature error term for Q2 (sharp quadrature estimate on Ω).
- It can be extended to Q^k with $k \geq 2$.
- This implementation is monotone thus satisfies Discrete Maximum Principle.

Numerical tests: Error at Gauss-Lobatto Points

	FEM using Gauss Lobatto Quadrature			
Mesh	l^2 error	order	l^∞ error	order
10×10	9.36E0	-	8.24E0	-
20×20	1.51E0	2.63	1.12E0	2.88
40×40	8.18E-2	4.21	8.35E-2	3.74
80×80	4.88E-3	4.07	8.54E-3	3.29
160×160	3.05E-4	4.00	1.09E-3	2.97
	FEM with Approximated Coefficients			
10×10	9.37E0	-	8.32E0	-
20×20	1.51E0	2.63	1.12E0	2.89
40×40	8.17E-2	4.21	7.36E-2	3.93
80×80	4.84E-3	4.08	5.00E-3	3.88
160×160	2.96E-4	4.03	3.38E-4	3.89
	Full FEM Scheme			
10×10	1.46E-1	-	4.31E-1	-
20×20	1.64E-2	3.16	6.55E-2	2.71
40×40	7.08E-4	4.53	3.42E-3	4.26
80×80	4.44E-5	4.06	4.84E-4	2.82
160×160	2.95E-6	3.85	7.96E-5	2.60

1D Constant Coefficient Case

The continuous P^2 FEM for $-u'' = f, u(0) = u(1) = 0$ is to solve $u_h \in V_0^h$ satisfying

$$\int_I u_h'(x)v_h'(x)dx = (f, v_h), \forall v_h \in V_0^h.$$

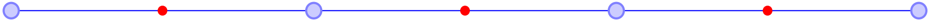
3-point Gauss Lobatto Quadrature

$$\int_I u_h'(x)v_h'(x)dx = \langle f, v_h\rangle_h, \forall v_h \in V_0^h.$$

Matrix-vector form $S\mathbf{u} = M\mathbf{f}$ or $M^{-1}S\mathbf{u} = \mathbf{f}$, which becomes

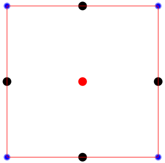
$$\text{midpoint} \qquad \frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2} = f_i$$

$$\text{endpoint} \qquad \frac{u_{i-2} - 8u_{i-1} + 14u_i - 8u_{i+1} + u_{i+2}}{4h^2} = f_i$$



Supraconvergence: second order in truncation error everywhere, but the L^2-norm error is third order.
Hao Li and X.Z. 2020c, JSC: this is a fourth order accurate (superconvergence) FD scheme for a 2D elliptic equation.

2D Constant Coefficient Case



$$Q2 \quad \text{cell center} \quad \begin{array}{ccc} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{array} \quad \text{edge center} \quad \begin{array}{ccccc} & & -1 & & \\ \frac{1}{4} & -2 & 2+\frac{7}{2} & -2 & \frac{1}{4} \\ & & -1 & & \end{array} \quad \text{vertex} \quad \begin{array}{ccccc} & & \frac{1}{4} & & \\ & & -2 & & \\ \frac{1}{4} & -2 & 7 & -2 & \frac{1}{4} \\ & & -2 & & \\ & & \frac{1}{4} & & \end{array}$$

Supraconvergence: second order in truncation error everywhere, but the L^2-norm error is third order.
Hao Li and X.Z. 2020c, JSC: this is a fourth order accurate (superconvergence) FD scheme for a 2D elliptic equation.

Q2 FD Scheme for Variable Coefficient

1D Equation $-(a(x)u')' = f$:

$$\text{midpoint}\,\frac{-(3a_{i-1}+a_{i+1})u_{i-1}+4(a_{i-1}+a_{i+1})u_i-(a_{i-1}+3a_{i+1})u_{i+1}}{4h^2}=f_i$$

$$\text{endpoint}\,\frac{(3a_{i-2}-4a_{i-1}+3a_i)u_{i-2}-(4a_{i-2}+12a_i)u_{i-1}+(a_{i-2}+4a_{i-1}+18a_i+4a_{i+1}+a_{i-2})u_i}{8h^2}$$

$$+\frac{-(12a_i+4a_{i+2})u_{i+1}+(3a_{i+2}-4a_{i+1}+3a_i)u_{i+2}}{8h^2}=f_i$$

2D Equation $-\nabla(a(x,y)\nabla u) = f$:

$$\text{cell center}\ \begin{array}{ccc} & * & \\ * & * & * \\ & * & \end{array}\quad \text{edge center}\ \begin{array}{ccccc} & & * & & \\ * & * & * & * & * \\ & & * & & \end{array}\quad \text{vertex}\ \begin{array}{ccccc} & & * & & \\ & & * & & \\ * & * & * & * & * \\ & & * & & \\ & & * & & \end{array}$$

M-Matrix

- If A is a nonsingular M-matrix, then $A^{-1} \geq 0$.
- Definition: a square matrix A that can be expressed in the form $A = sI - B$, where B has non-negative entries, and $s > \rho(B)$, the maximum of the moduli of the eigenvalues of B, is called an M-matrix.
- Sufficient but not necessary condition: if all the row sums of A are non-negative and at least one row sum is positive, then A is a a nonsingular M-matrix. Example:

$$\begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -1 & 2 & -1 & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}$$

- Sufficient and necessary condition: there exists a positive diagonal matrix D such that AD has all positive row sums. Example:

$$A = \begin{bmatrix} 10 & 0 & 0 \\ -10 & 2 & -10 \\ 0 & 0 & 10 \end{bmatrix}, D = \begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0.1 \end{bmatrix}, AD = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 4 & -1 \\ 0 & 0 & 1 \end{bmatrix}.$$

Monotonicity Implies Discrete Maximum Principle for Elliptic Equation

- Backward Euler for $u_t = u_{xx}$: $-\partial_{xx} u^{n+1} + \frac{1}{\Delta t} u^{n+1} = \frac{1}{\Delta t} u^n$.
- Maximum Principle: the solution of $-\nabla \cdot (a \nabla u) + cu = 0$ with $a > 0, c \geq 0$ in Ω with Dirichlet boundary condition g on $\partial\Omega$, then $\max_\Omega |u| \leq \max_{\partial\Omega} |g|$.
- Ciarlet 1970: Monotonicity Implies Discrete Maximum Principle.
- Second order centered difference is monotone for $-u'' = f, u(0) = u(1) = 0$:

$$-D_{xx} u = f, \quad -D_{xx} = \frac{1}{\Delta x^2} \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -1 & 2 & -1 & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}.$$

A matrix A is called monotone if its inverse has non-negative entries ($A^{-1} \geq 0$). $-D_{xx}$ is an M-matrix (diagonal entries are positive, off diagonal ones are non-positive, diagonally dominant and invertible) thus monotone.

Monotonicity is only a sufficient condition to achieve bound-preserving property. Advantage of using monotonicity: easier extension to more general equations and demanding applications.

Known Discrete Maximum Principle for High Order Schemes

2D constant coefficient case (the Laplacian operator):

- 9-point discrete Laplacian forms an M-matrix: Krylov and Kantorovitch, 1958; Collatz, 1960; Bramble and Hubbard 1963.
- 9-point scheme is one kind of Fourth Order Compact Finite Difference schemes, which form an M-matrix.
- Bramble and Hubbard 1964 4th order FD; matrix is not a M-matrix but can be factored a product of M-matrices: if $A = M_1 M_2$, then $A^{-1} = M_2^{-1} M_1^{-1} \geq 0$.
- Lorenz 1977 Lagrange P2 FEM on regular triangular mesh.
- Höhn and Mittelmann 1981 For P2, if angles are less than 90 degree, DMP holds only on equilateral triangulation.

Other results:

- Vejchodský and Šolín 2007 hp FEM (arbitrarily high order) in 1D (constant coef) satisfies DMP, via discrete Green's function.
- Vejchodský 2009 negative computational results for P3/P4/P5 in 2D.

Variable coefficient in 2D:

- Xu and Zikatanov 1999 P1 FEM on unstructured meshes, scalar coefficient.

M-matrix for Proving Monotonicity

M-matrix (diagonal entries are positive, off diagonal ones are non-positive, diagonally dominant and invertible) is the only tool to achieve monotonicity of a matrix A (A^{-1} has non-negative entries).

1. Schemes forming M-matrices:
    - $-\Delta u = f$: second order 5-point discrete Laplacian
    - $-\Delta u = f$: fourth order 9-point discrete Laplacian
    - $-\nabla(a\nabla u) = f$: P1 finite element on unstructured meshes.
2. If $A = M_1 M_2$, then $A^{-1} = M_2^{-1} M_1^{-1} \geq 0$:
    - $-\Delta u = f$: a fourth order FD scheme by Bramble and Hubbard 1964.
3. Lorenz 1977: for $A^{-1} \geq 0$, it suffices to show $A \leq ML$ where
    - M is an M-matrix.
    - Off-diagonal entries of L are negative and the sparsity pattern is the same as the one for negative entries in A.

Lorenz's Sufficient Condition for Monotonicity

Jens Lorenz, Zur inversmonotonie diskreter probleme, Numerische Mathematik (1977).

Assume diagonal entries of A are positive, and A becomes an M-matrix if setting positive off-diagonal entries to zero (example: high order schemes). Then for $A^{-1} \geq 0$, it suffices to show $A \leq ML$ where

- M is an M-matrix.
- Off-diagonal entries of L are negative and the sparsity pattern is the same as the one for negative entries in A.

1. Split A into three parts: diagonal, positive off-dial entries, negative off-diag entries $A = D + O^+ + O^-$.
2. Split negative entries: $O^- = Z + S$, $Z \leq 0$, $S \leq 0$ satisfying
   - $O^+ \leq ZD^{-1}S$ (only need to check positive entries in O^+ since $ZD^{-1}S \geq 0$)
   - S has the same sparsity pattern as O^-.
3. $A = D + Z + S + O^+ \leq D + Z + S + ZD^{-1}S = (D+Z)(I + D^{-1}S) = ML$.

1D Constant Coefficient Case

The continuous P^2 FEM for $-u'' = f, u(0) = u(1) = 0$ is to solve $u_h \in V_0^h$ satisfying

$$\int_I u_h'(x)v_h'(x)dx = (f, v_h), \forall v_h \in V_0^h.$$

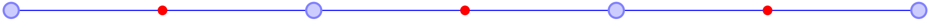
3-point Gauss Lobatto Quadrature

$$\int_I u_h'(x)v_h'(x)dx = \langle f, v_h\rangle_h, \forall v_h \in V_0^h.$$

Matrix-vector form $S\mathbf{u} = M\mathbf{f}$ or $M^{-1}S\mathbf{u} = \mathbf{f}$, which becomes

$$\text{midpoint} \qquad \frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2} = f_i$$

$$\text{endpoint} \qquad \frac{u_{i-2} - 8u_{i-1} + 14u_i - 8u_{i+1} + u_{i+2}}{4h^2} = f_i$$



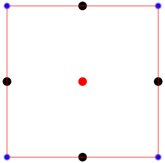
The matrix vector form is $\frac{1}{h^2}A\mathbf{u} = \mathbf{f}$ where

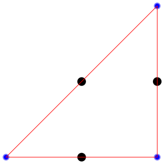
$$A = \begin{pmatrix} 2 & -1 & & & & & \\ -2 & \frac{7}{2} & -2 & \frac{1}{4} & & & \\ & -1 & 2 & -1 & & & \\ & \frac{1}{4} & -2 & \frac{7}{2} & -2 & \frac{1}{4} & \\ & & & -1 & 2 & -1 & \\ & & & \frac{1}{4} & -2 & \frac{7}{2} & -2 \\ & & & & & -1 & 2 \end{pmatrix}$$

$$A = M_1 M_2 = \begin{pmatrix} 1 & -\frac{1}{2} & & & & & \\ -1 & \frac{5}{2} & -1 & -\frac{1}{4} & & & \\ & -\frac{1}{2} & 1 & -\frac{1}{2} & & & \\ & -\frac{1}{4} & -1 & \frac{5}{2} & -1 & -\frac{1}{4} & \\ & & & -\frac{1}{2} & 1 & -\frac{1}{2} & \\ & & & -\frac{1}{4} & -1 & \frac{5}{2} & -1 \\ & & & & & -\frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} 2 & -\frac{1}{2} & & & & & \\ & 1 & & & & & \\ & -\frac{1}{2} & 2 & -\frac{1}{2} & & & \\ & & & 1 & & & \\ & & & -\frac{1}{2} & 2 & -\frac{1}{2} & \\ & & & & & 1 & \\ & & & & & -\frac{1}{2} & 2 \end{pmatrix}$$

- No geometrical/physical meaning.
- Cannot be extended to variable coefficient.
- Extension for 2D Laplacian (both Dirichlet and Neumann b.c.) is possible.

2D Constant Coefficient Case





$$Q2 \quad \text{cell center} \quad \begin{array}{ccc} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{array} \quad \text{edge center} \quad \begin{array}{ccccc} & & -1 & & \\ \frac{1}{4} & -2 & 2+\frac{7}{2} & -2 & \frac{1}{4} \\ & & -1 & & \end{array} \quad \text{vertex} \quad \begin{array}{ccccc} & & \frac{1}{4} & & \\ & & -2 & & \\ \frac{1}{4} & -2 & 7 & -2 & \frac{1}{4} \\ & & -2 & & \\ & & \frac{1}{4} & & \end{array}$$

$$P2 \quad \text{edge center} \quad \begin{array}{ccc} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{array} \quad \text{vertex} \quad \begin{array}{ccccc} & & 1 & & \\ & & -4 & & \\ 1 & -4 & 12 & -4 & 1 \\ & & -4 & & \\ & & 1 & & \end{array}$$

Q2 FD Scheme for Variable Coefficient

1D Equation $-(a(x)u')' = f$:

$$\text{midpoint}\,\frac{-(3a_{i-1}+a_{i+1})u_{i-1}+4(a_{i-1}+a_{i+1})u_i-(a_{i-1}+3a_{i+1})u_{i+1}}{4h^2}=f_i$$

$$\text{endpoint}\,\frac{(3a_{i-2}-4a_{i-1}+3a_i)u_{i-2}-(4a_{i-2}+12a_i)u_{i-1}+(a_{i-2}+4a_{i-1}+18a_i+4a_{i+1}+a_{i-2})u_i}{8h^2}$$

$$+\frac{-(12a_i+4a_{i+2})u_{i+1}+(3a_{i+2}-4a_{i+1}+3a_i)u_{i+2}}{8h^2}=f_i$$

2D Equation $-\nabla(a(x,y)\nabla u)=f$:

$$\text{cell center}\;\begin{array}{ccc} & * & \\ * & * & * \\ & * & \end{array}\;\text{edge center}\;\begin{array}{ccccc} & & * & & \\ * & * & * & * & * \\ & & * & & \end{array}\;\text{vertex}\;\begin{array}{ccccc} & & * & & \\ & & * & & \\ * & * & * & * & * \\ & & * & & \\ & & * & & \end{array}$$

Monotonicity and Discrete Maximum Principle for FD Q2 FEM

Hao Li and X.Z., Numerische Mathematik (2020):

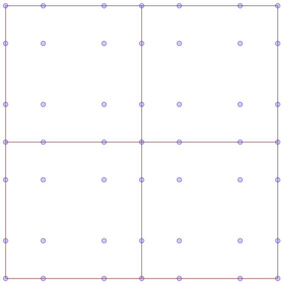
- For solving 1D (and 2D) variable coefficient Poisson equation $-\nabla \cdot (a \nabla u) = f$, Lorenz's condition can be achieved under reasonable mesh size constraint:
  1. $$h \max_e |\nabla a(x)| \leq \frac{1}{2} \min_e a(x).$$
  2. In 1D, if $a(x)$ is concave: then no constraint.
- For $-\nabla \cdot (a \nabla u^{n+1}) + \frac{1}{\Delta t} u^{n+1} = \frac{1}{\Delta t} u^n$: an additional lower bound on time step
  $$\frac{\Delta t}{h^2} \geq \frac{1}{5 \min a(x)}.$$

Catch:

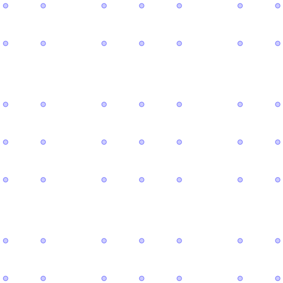
- DMP might be false if using more accurate quadrature.
- Lorenz's condition becomes tractable to verify for FD implementation.

Hao Li and X.Z., 2020c, JSC: FD Q2 FEM is a fourth order accurate scheme.

FD Q3 FEM for Laplacian



(o) Gauss-Lobatto Quadrature points and a finite element mesh.



(p) The corresponding finite difference grid.

Logan Cross and X.Z., ongoing: construct intermediate matrices such that Lorenz's condition can be applied recursively

$$
\begin{aligned}
&A_1 \le M_1 L_1 \Rightarrow A_1 = M_1 M_2 \Rightarrow A_2 \le M_1 M_2 L_2 \Rightarrow A_2 = M_1 M_2 M_3 \\
&\Rightarrow A \le M_1 M_2 M_3 L_3 \Rightarrow A = M_1 M_2 M_3 M_4 \Rightarrow A^{-1} \ge 0.
\end{aligned}
$$

Concluding Remarks

Summary:

- FD Q2 FEM is a fourth order monotone scheme for $\nabla \cdot (a\nabla u)$ with suitable mesh constraints.
- Superconvergence of the approximated coefficient (replace $\nabla(\mathbf{a}\nabla u)$ by $\nabla(\mathbf{a}_I\nabla u)$): for elliptic problems, a 4th order scheme can be obtained by using a 3rd order accurate coefficient.
  Possible extensions: wave equation, parabolic equations, Helmholtz equation...
- References on arXiv:
  - Hao Li and X.Z. 2020b JSC: superconvergence of the approximated coefficients for Q^k.
  - Hao Li and X.Z. 2020c JSC: superconvergence of the FD Q2 FEM.
  - Hao Li and X.Z. 2020a Numerische Mathematik : DMP for FD Q2 FEM.

Ongoing efforts on generalizations/applications:

- Logan Cross and X.Z.: $Q2$ on quasi-uniform grid for Laplacian.
- Logan Cross and X.Z.: $Q3$ on uniform grid for Laplacian.
- J. Shen and X.Z.: Maximum principle for implicitly solving diffusion in phase field equations (Allen-Cahn).
- J. Hu and X.Z.: Positivity and entropy decay of solving linear kinetic Fokker Planck equation.
- Unconditional stability in solving compressible Naiver-Stokes equations.

Wide open problem: Unconditionally stable high order implicit time solver.

Possible Applications

A fourth order accurate spatial upgrade: any positivity preserving method by second order FD or P1 FEM element can be extended to FD implementation of Q2 FEM.

- Backward Euler+second order FD for $u_t = \nabla(a\nabla u)$.
- Positivity for implicitly solving diffusion in phase field equations (Allen-Cahn):
  - Shen, Tang and Yang 2016: second order centered difference.
  - J. Xu, Li, Wu, and Bousquet 2018: P1 FEM.
- Positivity and entropy decay of solving linear kinetic Fokker Planck equation:
  - R. Bailo, J. Carrillo, and J. Hu: Backward Euler+second order FD
- Conservative Positivity-Preserving Methods for Compressible Navier-Stokes:
  - X.Z. 2017: fully explicit arbitrarily high order DG on unstructured meshes, general model of stress tensor and heat flux, but $\Delta t = \mathcal{O}(Re\Delta x^2)$ thus only suitable for high Reynolds number flows.
  - D. Grapsas, R. Herbin, W. Kheriji, J.-C. Latché, 2015: second order implicit scheme unconditionally stable, but only for simplified dimensionless form of the compressible NS system (Laplacian on the internal energy)

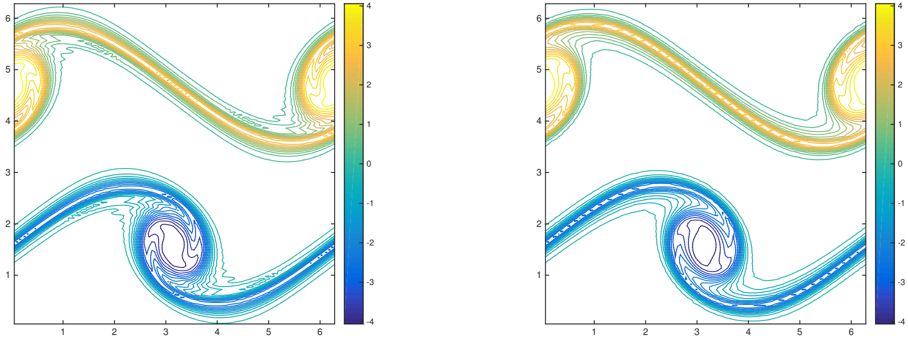
SSP type Runge-Kutta (convex combination of backward Euler) has an additional time step constraint $\Delta t = \mathcal{O}(\Delta x^2)$.

Allen Cahn equation with a passive convection term

Finite difference schemes on a 239 × 239 mesh. Time discretization is backward Euler. The solution on the left is wrong. Higher order time discretization does not improve the error for second order finite difference on such a relative coarse mesh.

2D incompressible Navier Stokes in vorticity form

Double shear layer: finite difference with backward Euler on a 120 × 120 grid. Viscosity coefficient $\mu = 0.001$.



(q) Second order difference on a 120 × 120 grid

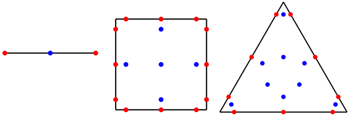
(r) Fourth order difference on a 120 × 120 grid

Figure: The fourth order scheme is obviously superior, even though only first order time discretization is used and sharp gradient is involved.

Take Home Message: Monotonicity in High Order Schemes

Part I: Weak monotonicity in explicit high order schemes for convection diffusion problems: design a limiter to control lower bound of any concave or quasi-concave quantities.

- Example: internal energy $e = E - \frac{1}{2}\rho\|\mathbf{u}\|^2$ is concave, entropy $S = \log\frac{p}{\rho^\gamma}$ is quasi-concave is gas dynamics.
- Advantage: easy limiter for complicated system/geometry, rigorous justification of accuracy.



Part II: some finite difference schemes can be perceived as finite volume schemes.

Part III: Monotonicity (inverse positivity $L^{-1} \geq 0$) for solving linear diffusion implicitly.

- Hao Li and X.Z. Numerische Mathematik (2020): FD Q2 FEM is a fourth order monotone scheme for $\nabla \cdot (a\nabla u)$ with suitable mesh constraints.