INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

New two-dimensional slope limiters for discontinuous Galerkin methods on arbitrary meshes

Hussein Hoteit — Philippe Ackerer — Robert Mosé —

Jocelyne Erhel — Bernard Philippe

N° 4491

Juillet 2002

THÈME 4

ISSN 0249-6399 ISRN INRIA/RR--4491--FR+ENG

INRIA
RENNES

New two-dimensional slope limiters for discontinuous Galerkin methods on arbitrary meshes

Hussein Hoteit , Philippe Ackerer* , Robert Mosé† ,
Jocelyne Erhel, Bernard Philippe

Thème 4 — Simulation et optimisation
de systèmes complexes
Projet Aladin

Rapport de recherche n° 4491 — Juillet 2002 — 37 pages

Abstract: In this paper we introduce an extension of Van Leer's slope limiter for two-dimensional Discontinuous Galerkin (DG) methods on arbitrary unstructured quadrangular or triangular grids. The aim is to construct a non-oscillatory shock capturing DG method for the approximation of hyperbolic conservative laws without adding excessive numerical dispersion. Unlike some splitting techniques that are limited to linear approximations on rectangular grids, in this work, the solution is approximated by means of piecewise quadratic functions. The main idea of this new reconstructing and limiting technique follows a well-known approach where local maximum principle regions are defined by enforcing some constraints on the reconstruction of the solution. Numerical comparisons with some existing slope limiters on structured as well as on unstructured meshes show a superior accuracy of the proposed slope limiters.

Key-words: hyperbolic conservative laws, discontinuous Galerkin methods, slope limiters, upwind schemes.

* Institut de Mécanique des Fluides et des solides, Univ. Louis Pasteur de Strasbourg, CNRS/MUR 7507, 2 rue Boussingault, 67000 Strasbourg.
† Ecole Nationale du Génie de l'Eau et de l'Environnement 1, Quai Koch, 67070 Strasbourg.

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 Rennes Cedex (France)
Téléphone : +33 2 99 84 71 00 —Télécopie : +33 2 99 84 71 71

Nouveaux limiteurs de pente bidimensionnels pour la méthode des élément finis discontinus de Galerkin sur des maillage arbitraires

Résumé : Dans ce travail nous présentons une extension du limiteur de pente de Van Leer pour la méthode des Eléments Finis Discontinus de Galerkin (EFD) sur des maillages quadrangulaires ou triangulaires non structurés. Le but est de construire une méthode des EFD non-oscillante qui puisse capturer les chocs en approchant les lois de conservation sans ajouter de la dispersion numérique excessive. A la différence de certaines techniques de séparation qui se limitent aux approximations linéaires sur les maillages rectangulaires, dans ce travail, la solution est approchée par des fonctions quadratiques par morceaux. L'idée principale de cette nouvelle technique de limitation suit une approche bien connue où un principe du maximum local est défini à partir de quelques contraintes sur la reconstruction de la solution. Des comparaisons numériques avec quelques limiteurs de pente existants sur des maillages structurés et non structurés montrent la supériorité des limiteurs de pente proposés.

Mots-clés : lois de conservation hyperbolique, éléments finis discontinus de Galerkin, limiteur de pente, schéma amont.

1 Introduction

The success of DG methods in approximating various physical problems notably hyperbolic systems of conservative laws has attracted the attention to explore the benefits of this approach. One favorable property of DG methods is that they conserve mass at the element level in a finite element frame work. Consequently, they inherit the flexibility of finite elements in handling complicated geometries. Furthermore, the particular approximation space of these methods, where continuity across inter-element boundaries is needless, allows a simple treatment of non homogeneous finite element geometries as well as different degree of approximating polynomials. It is known that when using constant cell approximations the numerical diffusion due to upwinding is big enough to keep the scheme stable. However, by using higher order approximation spaces the scheme produces non physical oscillations near shocks. In such a case, the use of an appropriate slope limiter is crucial to ensure the stability of the method.

In one dimensional space, discontinuous finite elements can be interpreted as a generalization of high order Godunov finite differences [12, 24, 25, 26]. Such high resolution schemes are usually stabilized using some form of TVD (Total Variation Diminishing) limiters (see, e.g., [23, 17]) so that spurious oscillations can be avoided without destroying the high order accuracy of the schemes. One commonly used technique is the Van Leer's MUSCL (Monotonic Upstream Centered Schemes for Conservative Laws) slope limiter [26]. In the works of Cockburn and Shu [21, 7, 8], this slope limiter is extended to the so-called generalized slope limiter where a $(k+1)$-th order of accuracy is achieved in smooth regions by using DG method with polynomials of degree k for the spatial discretization and a special $(k+1)$-th order explicit Runge-Kutta method for temporal discretization. The generalized slope limiter does not totally smear oscillations near shocks as a way to prevent spoiling the scheme accuracy in smooth regions. Thus, the resulting scheme is no more TVD, however it satisfies a TVB (Total Dimension Bounded) property.

In multi-dimensional spaces, DG methods are still facing difficulties to attain the same degree of accuracy as in the one-dimensional case, specially on unstructured meshes. The troublesome part is the construction of appropriate multi-dimensional slope limiters that preserve the accuracy of the scheme. Nevertheless, it is proved that any scheme combined with a slope limiting operator that enforces a TVD condition is at most first order accurate [13]. Consequently, a great deal of effort has been oriented for the construction of genuinely multi-dimensional slope limiters that

can eliminate unphysical oscillations without adding excessive numerical viscosity. One simple approach in the case of rectangular grids is to use the DG method with linear polynomials (P^1) for the space discretization instead of quadratic ones (Q^1) [11]. This approach can be considered as a dimensional splitting technique [23]. In this case, the slope limiting process can be carried out by applying a one-dimensional slope limiter sequentially in the x- and y-directions.

In our work, we concentrate on a genuinely multi-dimensional slope limiter in the sense that it does not require any operator splitting. This slope limiting operator was introduced by Chavent and Jaffré [4] as a generalization of Van Leer's MUSCL limiter [26]. It can be applied in a geometric manner so that slopes are limited in such a way that each sub-reconstruction lies between the cell averages of its neighbors. In the one dimensional case, Gowda and Jaffré have analyzed this limiter and proved the stability of the DG method with the TVD property [10, 11]. Nevertheless, we have found that the proposed extension of this limiter to the multi-dimensional case does not give satisfactory results. We have detected some cases for both triangular and rectangular discretizations where the limiting operator fails to completely eliminate undershots and overshots. The origin of this drawback is due to the fact that limiting slopes by using the nodal values of the solution does not prohibit unphysical values at the midpoints of the cell edges. As a result, this approach does not satisfy a local maximum principle. This paper proposes a remedy whereby this limiting technique can be improved by taking more account of the averages of the cell edges.

For triangular elements, piecewise linear approximations are used with degrees of freedom at the grid vertices. Our limiting process intends to reconstruct the solution first at the midpoints of the cell edges by preventing local extrema then at the cell vertices by using the midpoint reconstructions. This is less restrictive than reconstructing directly at the cell vertices. On the other hand, we have found that by taking the degrees of freedom at the midpoints of the grid edges, the scheme leads to excessive smearing.

For rectangular elements, the solution is approximated by using piecewise quadratic function where the degrees of freedom are indeed at the grid vertices. We use similar techniques for the reconstructions at the midpoints of the edges within each cell. Unfortunately, the information at the midpoints of the edges is not sufficient to give a unique reconstruction at the cell vertices. Consequently, we append supplemen-

tary constraints in order to overcome the singularity of the system.

The discontinuous Galerkin finite element method for scalar, linear conservation laws is reviewed in the next section. In section three, we present the slope limiter introduced by Chavent and Jaffré in one and higher dimensional spaces. We give a simple numerical test where this limiter fails to eliminate all oscillations. Section four is devoted to describe our modified slope limiter for unstructured rectangular and triangular grids. In section five, we briefly review some existing slope limiters, in particular those introduced by Cockburn and Shu. Finally, before ending with a conclusion in section seven, we give, in section six, some critical comparisons between the described reconstruction techniques by using several numerical experiments.

2 Discontinuous Galerkin finite element method

The ultimate goal of this work is to check out the reconstruction techniques for DG methods. Thus, for the sake of brevity, we restrict our attention to two-dimensional, linear, scalar advection equations. The extension to three-dimensional general conservation laws is an ongoing work.

Hence, we consider the hyperbolic-type equation of the form

$$\frac{\partial u}{\partial t} + \nabla . f(u) = 0 \text{ in } \Omega \times (0,T), \tag{1}$$

with the initial conditions

$$u(x,0) = u^0(x) \text{ in } \Omega, \tag{2}$$

and appropriate boundary conditions. Here $f(u) = u\,\beta$ where $u = u(x,t)$ is a scalar unknown representing a concentration for example, $\beta = (\beta_1(x), \cdots, \beta_d(x))$, $(d = 1, 2)$ is a given vector field, $\Omega \subset \mathbb{R}^d$ and $(0,T)$ is a given time interval.

2.1 Space integration

In this presentation of the DG method, some materials are drawn from these works [4, 6, 18, 11, 9]. The polygonal domain Ω is discretized into a mesh $\mathcal{T}_h$ consisting of quadrilaterals or triangles where h refers to the maximal element diameter. We also denote by $\mathcal{N}_K$ the number of vertices of the discretized element K.

The discontinuous Galerkin method is based on using the following discontinuous finite element space :

$$V_h = \{v \in L^\infty(\Omega) : v_h|_K \in V(K),\ \forall K \in \mathcal{T}_h\},$$

where $V(K)$ is the space of linear P^1 (resp. quadratic Q^1) polynomials if K is a triangle (resp. quadrilateral). In this work, we are restricted to polynomials of degree one.

In order to define the upwind technique [17] we need to split the boundary ∂K of a discretized element K into an inflow part ∂K^{in} and an outflow part ∂K^{out} defined by

$$\partial K^{in} = \{x \in \partial K : n(x).\beta \geq 0\},$$
$$\partial K^{out} = \{x \in \partial K : n(x).\beta < 0\},$$

where $n(x)$ denotes the unit outward normal to ∂K (see Fig. 1).

Let E be a common edge between any two adjacent elements K and K'. Since discontinuity for any function $v \in V_h$ is allowed across interelement boundaries, we need to define the jump discontinuity of v across E. We introduce the notations v^{in} and v^{out} to denote respectively the inner and the outer values of v over E with respect to K, that is,

$$\begin{aligned} v^{in}(x) &= \lim_{\epsilon \to 0^+} v(x + \epsilon\beta), \quad x \in \partial K, \\ v^{out}(x) &= \lim_{\epsilon \to 0^-} v(x + \epsilon\beta), \quad x \in \partial K. \end{aligned}$$

The formulation obtained by using the discontinuous Galerkin method is formulated by multiplying equation (1) by a sufficiently smooth test function v and by integrating by parts over an element $K \in \mathcal{T}_h$

$$\int_K \frac{\partial u}{\partial t} v \, dx + \int_K f(u).\nabla v \, dx - \int_{\partial K} v f(u).n \, d\ell = 0. \tag{3}$$

Then, we replace u by the approximate solution u_h which can be expressed as follows :

$$u_h(x,t) \equiv u_h(x,t)|_K = \sum_{j=1}^{\mathcal{N}_K} u_{K,j}(t)\varphi_{K,j}(x) = \sum_{j=1}^{\mathcal{N}_K} u^{in}_{K,j}(t)\varphi_{K,j}(x),$$

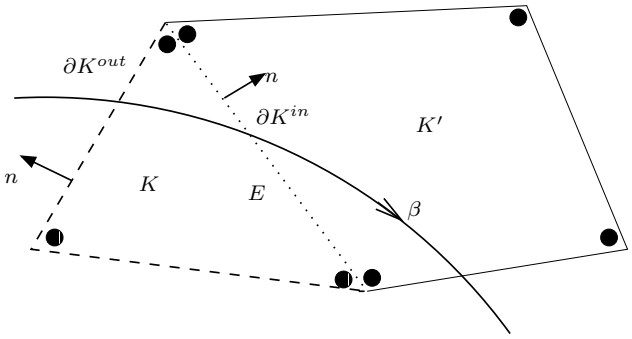


Figure 1: Inflow and outflow boundaries with the local degrees of freedom.

where $\varphi_{K,j}$ are some test functions in V_h that form a basis for the local approximation space $V(K)$. The standard finite element shape functions may be chosen.

Due to the discontinuity of u across ∂K, the flux function $f(u)$ is approximated by solving a one-dimensional Riemann problem. In our case $f(u) = u\beta$ is a linear function, consequently the Riemann solver is evident (see, e.g., [23]), that is,

$$f(u) = \begin{cases} f(u^{in}) & \text{over} \quad \partial K^{in}, \\ f(u^{out}) & \text{over} \quad \partial K^{out}. \end{cases}$$

By replacing v successively by the test functions $\varphi_{K,i}, i = 1, \cdots, \mathcal{N}_K$, the weak formulation (3) takes the following form :

$\forall K \in \mathcal{T}_h$, we seek the approximation solution $u_h \equiv u_h|_K \in V_h$ with the initial data (2) such that,

$$\sum_{j=1}^{\mathcal{N}_K} \frac{du_{K,j}}{dt} \int_K \varphi_{K,i}\varphi_{K,j}\, dx = -\sum_{j=1}^{\mathcal{N}_K} \left(u_{K,j} \int_K \varphi_{K,j}\beta.\nabla\varphi_{K,i}\, dx \right.$$
$$\left. + u^{in}_{K,j} \int_{\partial K^{in}} \varphi_{K,i}\varphi_{K,j}\beta.n\, d\ell + u^{out}_{K,j} \int_{\partial K^{out}} \varphi_{K,i}\varphi_{K,j}\beta.n\, d\ell \right). \quad (4)$$

Note that the inner values of the functions $\varphi_{K,i}$ are taken in the integrals across the boundaries of K in Eq.4.

2.2 Time integration

The DG approximation leads to a system of $\mathcal{N}_K$ ordinary differential equations over each element $K \in \mathcal{T}_h$. After inverting the local mass matrix, which corresponds to the integrals on the left-hand side of Eq.(4), this system can be rewritten in matrix form as follows :

$$\frac{dU_K}{dt} = \mathcal{A}(U_K^{in}, U_K^{out}), \tag{5}$$

where U_K is a vector of dimension $\mathcal{N}_K$ containing the cell unknowns $u_{_{K,j}}$ and $\mathcal{A}$ represents the components of the right-hand side of Eq.4 multiplied by the inverse of the mass matrix. Indeed, both notations U_K and U_K^{in} are the same.
In order to approximate the systems (5), we subdivide the time interval $[0, T]$ into a finite number of sub-interval $[t^n, t^{n+1}]$. Let $\Delta t = t^{n+1} - t^n$ denote the time step. We specify the following schemes :

2.2.1 Forward Euler method

A simple approach is to use Euler forward time discretization scheme. However, Chavent and Cockburn [5] showed that without using a suitable slope limiter this scheme is unconditionally unstable. Thus, the reconstruction process is crucial in order to stabilize the scheme. As a result, the DG computation procedure is illustrated by the following two steps :

1. Calculation of $\widetilde{U}_K^{n+1}$ for given u_h^n as follows

$$\widetilde{U}_K^{n+1} = U_K^n + \Delta t \mathcal{A}(U_K^{in,n}, U_K^{out,n}), \qquad \forall K \in \mathcal{T}_h.$$

2. Reconstruction of the updated solution $\widetilde{U}_K^{n+1}$ by applying

$$U_K^{n+1} = \mathcal{L}(\widetilde{U}_K^{n+1}), \qquad \forall K \in \mathcal{T}_h.$$

where $\mathcal{L}$ denotes a slope limiting operator to be discussed in the next section.

2.2.2 Explicit Runge-Kutta method

A second order accuracy in time may be obtained by using an explicit Runge-Kutta method. The time-stepping algorithm reads in four steps as follows :

1. Compute an intermediate function $\widetilde{U}_K^{n+1/2}$ for given u_h^n,

$$\widetilde{U}_K^{n+1/2} = U_K^n + \frac{\Delta t}{2}\mathcal{A}(U_K^{in,n}, U_K^{out,n}), \qquad \forall K \in \mathcal{T}_h.$$

2. Apply the slope limiter operator, $U_K^{n+1/2} = \mathcal{L}(\widetilde{U}_K^{n+1/2}), \qquad \forall K \in \mathcal{T}_h.$

3. Compute $\widetilde{U}_K^{n+1}$ for given u_h^n and $u_h^{n+1/2}$,

$$\widetilde{U}_K^{n+1} = U_K^n + \Delta t\mathcal{A}(U_K^{in,n+1/2}, U_K^{out,n+1/2}), \qquad \forall K \in \mathcal{T}_h.$$

4. Apply the slope limiter operator, $U_K^{n+1} = \mathcal{L}(\widetilde{U}_K^{n+1}), \qquad \forall K \in \mathcal{T}_h.$

2.2.3 Simplified Runge-Kutta

Due to the expensive computing cost for Riemann solvers as well as slope limiting process, a simplified version of the above Runge-Kutta method was introduced. This schema is used in these works [11, 16, 19, 22, 2]. Thus, the following three steps algorithm can be used instead :

1. Compute an intermediate function $\widetilde{U}_K^{n+1/2}$ for given u_h^n,

$$\widetilde{U}_K^{n+1/2} = U_K^n + \frac{\Delta t}{2}\mathcal{A}(U_K^{in,n}, U_K^{in,n}), \qquad \forall K \in \mathcal{T}_h.$$

2. Compute $\widetilde{U}_K^{n+1}$ for given u_h^n and $u_h^{n+1/2}$,

$$\widetilde{U}_K^{n+1} = U_K^n + \Delta t\mathcal{A}(U_K^{in,n+1/2}, U_K^{out,n+1/2}), \qquad \forall K \in \mathcal{T}_h.$$

3. Apply the slope limiter operator, $U_K^{n+1} = \mathcal{L}(\widetilde{U}_K^{n+1}), \qquad \forall K \in \mathcal{T}_h.$

Note that in the first step, the intermediate functions are calculated by means of local interior values of u_h^n.

3 Data Reconstruction

In this section, we focus on the slope limiter introduced by Chavent and Jaffré [4]. This limiter can be interpreted as a generalization of Van Leer's MUSCL limiter [26]. The essential idea of this technique is to impose some local constraints in a

geometric manner so that the reconstructed solution satisfies an appropriate maximum principle. Before starting with the multi-dimensional case, we present the slope limiter with one variable in space.

3.1 One dimensional slope limiter

Let us now denote by $K_i =]x_{i-1/2}\,,\, x_{i+1/2}[$ the sub-intervals of the one-dimensional space discretization. The sought function u_h is approximated by means of piecewise linear functions. We denote by $\overline{u}_i$ the sliding average of u_h over K_i which is indeed the midpoint of the two boundary nodal values, that is,

$$\overline{u}_i = \frac{1}{|K_i|}\int_{K_i} u_h\, dx = \frac{1}{2}(u_{i-1/2} + u_{i+1/2}).$$

For a given function $\widetilde{u}_h \in V_h$, set $u_h = \mathcal{L}(\widetilde{u}_h) \in V_h$. In order to reconstruct the solution $\widetilde{u}_h$ over an element K_i, only information of $\widetilde{u}_h$ over the adjacent elements K_{i-1} and K_{i+1} is needed. Thus, the problem reads as :

Find $U_i = (u_{i-1/2}, u_{i+1/2})$ such that the following conditions are satisfied :

1. Conservation of mass :

$$\overline{u}_i = \frac{1}{2}(u_{i-1/2} + u_{i+1/2}) = \frac{1}{2}(\widetilde{u}_{i-1/2} + \widetilde{u}_{i+1/2}).$$

2. Avoid creating local extremum $\forall \alpha \in (0,1)$:

$$\begin{array}{lll} (1-\alpha)\overline{u}_i + \alpha\min(\overline{u}_{i-1}, \overline{u}_i) & \leq u_{i-1/2} \leq & (1-\alpha)\overline{u}_i + \alpha\max(\overline{u}_{i-1}, \overline{u}_i) \\ (1-\alpha)\overline{u}_i + \alpha\min(\overline{u}_i, \overline{u}_{i+1}) & \leq u_{i+1/2} \leq & (1-\alpha)\overline{u}_i + \alpha\max(\overline{u}_i, \overline{u}_{i+1}) \end{array}$$

3. Minimum modification of $\widetilde{u}_h$: U_i is chosen as close as possible to $\widetilde{U}_i$ with respect to the L^2 norm, i.e.,

$$\|U_i - \widetilde{U}_i\|_2 \text{ is minimal.}$$

The above problem can also be rewritten using another more familiar form :

$$\begin{array}{rcl} u_{i-1/2} & = & \overline{u}_i - \mathcal{M}\left(\overline{u}_i - \widetilde{u}_{i-1/2}, \alpha(\overline{u}_i - \overline{u}_{i-1}), \alpha(\overline{u}_{i+1} - \overline{u}_i)\right), \\ u_{i+1/2} & = & \overline{u}_i + \mathcal{M}\left(\widetilde{u}_{i+1/2} - \overline{u}_i, \alpha(\overline{u}_i - \overline{u}_{i-1}), \alpha(\overline{u}_{i+1} - \overline{u}_i)\right). \end{array}$$

where $\mathcal{M}$ is the well known *minmod* function [15],

$$\mathcal{M}(a_1, a_2, a_3) = \begin{cases} s \min\limits_{1 \le i \le 3} |a_i| \text{ if } s = sign(a_1) = sign(a_2) = sign(a_3), \\ \\ 0 \qquad\qquad \text{otherwise.} \end{cases} \tag{6}$$

Note that $\mathcal{M}$ needs to be applied only once since

$$\begin{aligned} &\mathcal{M}\left(\overline{u}_i - \widetilde{u}_{i-1/2}, \alpha(\overline{u}_i - \overline{u}_{i-1}), \alpha(\overline{u}_{i+1} - \overline{u}_i)\right) \\ &= \quad \mathcal{M}\left(\widetilde{u}_{i+1/2} - \overline{u}_i, \alpha(\overline{u}_i - \overline{u}_{i-1}), \alpha(\overline{u}_{i+1} - \overline{u}_i)\right). \end{aligned}$$

The parameter α controls the degree of restriction of slopes, that is, the added

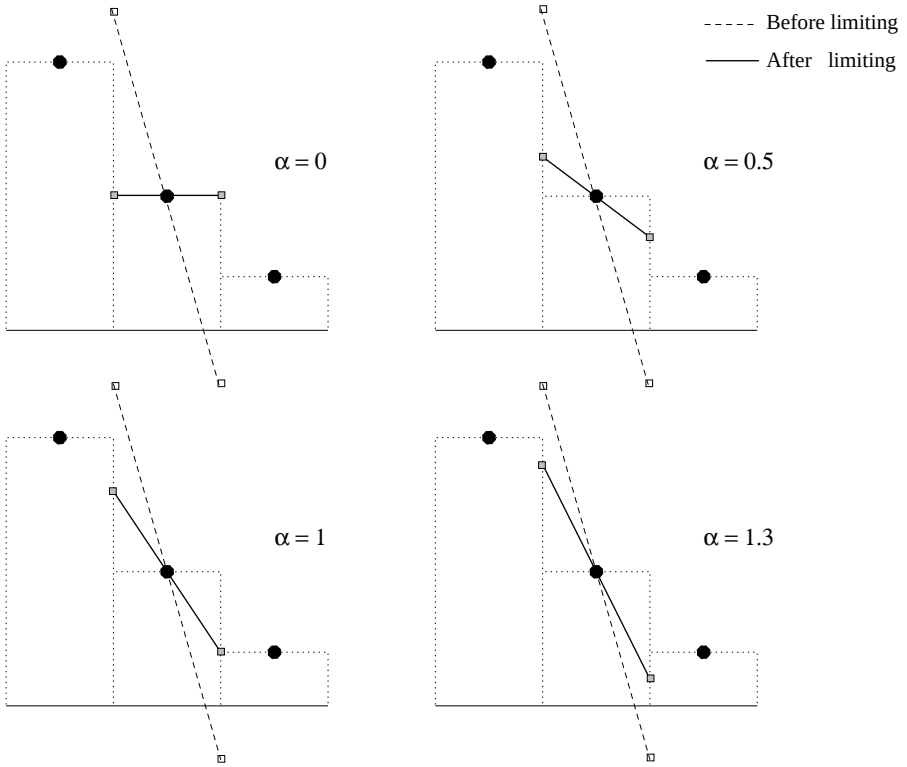


Figure 2: Function of the slope limiter with different values for α.

numerical viscosity (see Fig. 2). By choosing specific values for α, some well known slope limiters are obtained.

- For $\alpha = 0$, the slope limiter enforces constant piecewise approximations. Therefore, the scheme boils down to first order Godunov finite difference method [12].
- For $\alpha = 1/2$, we obtain the slope limiter of the MUSCL schemes of Van Leer [26].
- For $\alpha = 1$, we find a less restrictive limiter due to Osher [20].

By using Harten's TVD conditions [14], Gowda and Jaffré [11] proved the stability of the DG method combined with this slope limiter for $\alpha \in [0, 1]$. However, by taking α slightly greater than one, it is found that this slope limiter for smooth initial conditions behaves in a similar manner as the TVB generalized slope limiter introduced by Cockburn and Shu [7, 8].

3.2 Multi-dimensional slope limiter

The extension of the slope limiter to the multi-dimensional case is formulated in such a way that in each cell K each state variable at a vertex A_i lies between the cell averages of all neighboring elements containing A_i as a vertex. For any $K \in \mathcal{T}_h$, we introduce the following notations :

$$\begin{aligned} T(A) &= \{K \in \mathcal{T}_h |\ A \text{ is a vertex of } K\}, \\ U_K &= (u_{K,i})_{i=1,\cdots,\mathcal{N}_K}, \\ \overline{u}_K &= \frac{1}{|K|}\int_K u_h\, dx, \\ \overline{u}_{\min,i} &= \min_{K \in T(A_i)} \overline{u}_K, \\ \overline{u}_{\max,i} &= \max_{K \in T(A_i)} \overline{u}_K. \end{aligned}$$

The slope limiting process seeks $U_K \in V(K),\ \forall K \in \mathcal{T}_h$, the solution of the following least squares problem :

$$\min_W \|W - \widetilde{U}_K\|_2, \quad \text{subject to the linear constraints :} \tag{7}$$

$$\overline{w} = \frac{1}{\mathcal{N}_K}\sum_{j=1}^{\mathcal{N}_K} w_i = \overline{u}_K,$$

$$\forall \alpha \in [0,1],\ (1-\alpha)\overline{u}_K + \alpha\overline{u}_{\min,i} \le w_i \le (1-\alpha)\overline{u}_K + \alpha\overline{u}_{\max,i},\ \ i = 1, \cdots, \mathcal{N}_K.$$

It is easy to check that this minimization problem has a unique solution [4]. See appendix A for another robust algorithm.

We have found that this slope limiter sometimes fails to smear completely the spurious oscillations. Its weak point is that it does not prevent creating new extrema at the midpoints of the grid edges. In other words, it is possible to obtain a value of the state average over an edge E which is beyond the cell averages of the two adjacent grid elements having E as a common edge. As a result, we could have regions where a local maximum principle is violated.

3.3 Numerical test

In order to clarify the drawback of this slope limiter, we consider a very simple numerical test. Let $\Omega = (0, 10) \times (0, 10)$ be the computational domain and $\beta = (1, 0)$ be the velocity field. Two uniform grids of rectangles and triangles are considered with space steps $\Delta x = \Delta y = 1$ (see Fig.3). The scalar convection equation (1) is considered with zero initial conditions and a non smooth step Dirichlet boundary conditions on the left-hand side of the domain (Fig.3). Even though, this problem is physically one-dimensional, the above described multi-dimensional slope limiter is used. In figures 4 and 5, the cell average values of the solutions obtained by

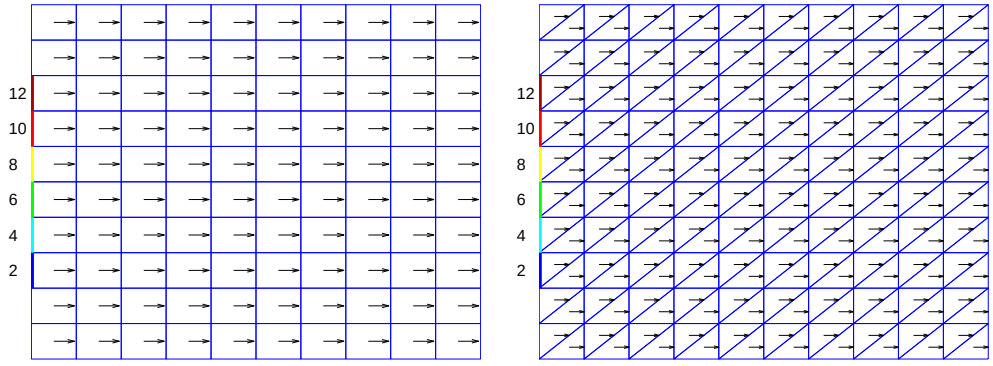
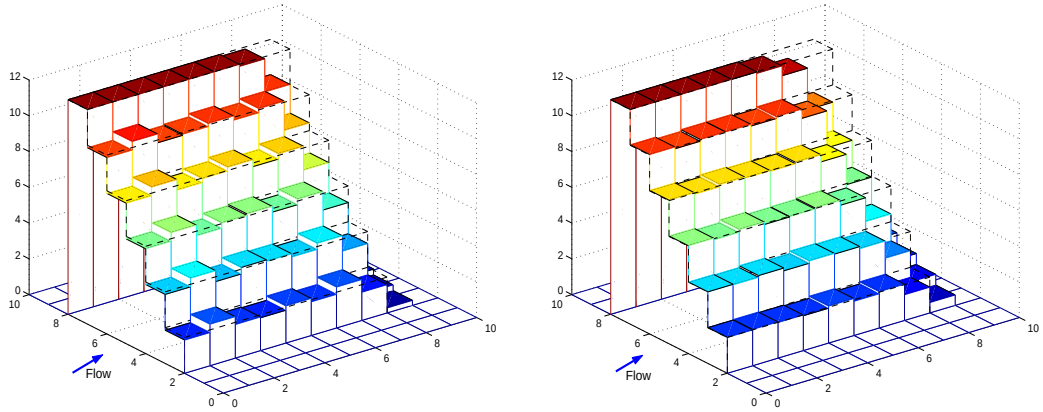


Figure 3: Uniform grids with Dirichlet boundary conditions at the left-hand side of the domain.

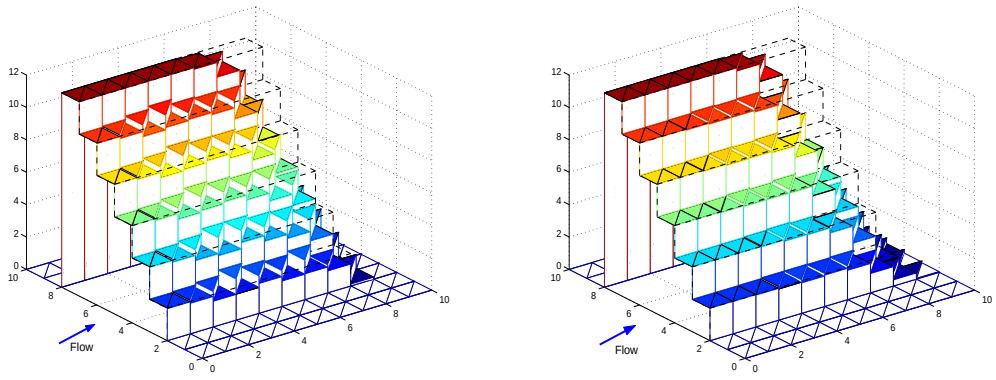
the DG method are presented without any graphical smoothing. The dashed lines represent the profile of the exact solution. It is clear that the slope limiter for both triangular and rectangular grids does not completely eliminate the non physical

oscillations. Decreasing the value of α will smear oscillations, however the scheme becomes more diffusive. The simplified Runge-Kutta method is used for the time integration. The time step is chosen so that the $CFL = \beta_x \frac{\Delta t}{\Delta x}$ condition is equal to 0.9 for rectangular grid and 0.4 for triangular grid. It should be noted that all temporal schemes introduced in the previous section produce similar results.



$\alpha = 1$ $\alpha = 0.1$

Figure 4: DG solutions on a rectangular grid at $T = 8$ with different values of α.



$\alpha = 1$ $\alpha = 0.1$

Figure 5: DG solutions on a triangular grid at $T = 8$ with different values of α.

4 Modified slope limiter

A remedy for this difficulty is possible by preventing the reconstruction to produce any new extrema at the midpoints of edges within each cell. This approach has an important physical property since it limits the interelement numerical fluxes rather than the function values at the grid vertices. However, the previous slope limiter does not satisfy this property. In the sequence, we introduce two new slope limiters for rectangular and triangular unstructured grids that satisfy this property.

4.1 Slope limiting for rectangular elements

Up to our knowledge not many slope limiters are available in literature for quadrangular elements. The slope limiter proposed by Cockburn and Shu in [9] for rectangular grids is essentially designed for linear cell approximations (P^1). Our slope limiter is related in some way to that limiter, however, we use piecewise quadratic polynomials (Q^1) to approximate the solution.
In order to formulate our slope limiter, we choose an arbitrary quadrangular element K_0 surrounded by its neighbors $K_i,\ i = 1, \cdots, 4$, as illustrated in Fig. 6. We denote by $A_{i,j}$ the midpoints of the edge $[A_i, A_j]$ and by $u_{i,j}$ the state average over the edge $[A_i, A_j]$. Indeed, $u_{i,j}$ is the midpoint of $[u_i, u_j]$.

Thus, unphysical oscillations at the midpoints $A_{i,j}$ can be avoided by enforcing the edge average $u_{i,j}$ to be within the averages of cells containing $[A_i, A_j]$ as a common edge, that is,

$$\begin{array}{lll} (1-\alpha)\overline{u}_{K_0} + \alpha \min(\overline{u}_{K_1}, \overline{u}_{K_0}) & \leq u_{1,2} \leq & (1-\alpha)\overline{u}_{K_0} + \alpha \max(\overline{u}_{K_1}, \overline{u}_{K_0}), \\ (1-\alpha)\overline{u}_{K_0} + \alpha \min(\overline{u}_{K_0}, \overline{u}_{K_3}) & \leq u_{3,4} \leq & (1-\alpha)\overline{u}_{K_0} + \alpha \max(\overline{u}_{K_0}, \overline{u}_{K_3}), \end{array} \tag{8}$$

$$\begin{array}{lll} (1-\alpha)\overline{u}_{K_0} + \alpha \min(\overline{u}_{K_2}, \overline{u}_{K_0}) & \leq u_{2,3} \leq & (1-\alpha)\overline{u}_{K_0} + \alpha \max(\overline{u}_{K_2}, \overline{u}_{K_0}), \\ (1-\alpha)\overline{u}_{K_0} + \alpha \min(\overline{u}_{K_0}, \overline{u}_{K_4}) & \leq u_{4,1} \leq & (1-\alpha)\overline{u}_{K_0} + \alpha \max(\overline{u}_{K_0}, \overline{u}_{K_4}). \end{array} \tag{9}$$

Therefore, an evident choice of the slope limiter is to add these constraints to system (11). Unfortunately, the resolution of the resulting minimization problem is computationally very expensive even when the constraints at the cell vertices are ignored. A key solution to overcome this difficulty is to replace the inequality constraints given in (8) and (9) by equality constraints. This can be done by adapting a dimension splitting technique that is illustrated by the following stages.

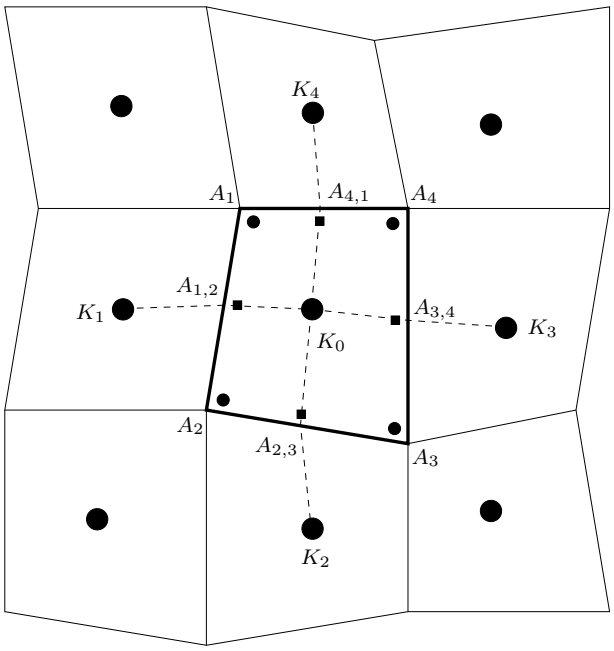


Figure 6: Illustration of limiting for quadrilateral elements.

1. We first reconstruct the state averages at the midpoints of the edges by using a direction splitting. Let us start in the x-direction, for example. Here, only information about the neighbor cells K_1 and K_3 is needed. Due to mass balance, we should have :

$$\frac{1}{2}(u_{1,2} + u_{3,4}) = \overline{u}_{K_0}.$$

By combining the local constraints subjected to the edge averages $u_{1,2}$ and $u_{3,4}$ (8), the resulting problem is thus reduced to a one-dimensional linear reconstruction. Consequently, we apply the one-dimensional slope limiter $\mathcal{M}$ previously discussed.

$$\begin{aligned} u_{1,2} &= \overline{u}_{K_0} - \mathcal{M}\left(\overline{u}_{K_0} - \widetilde{u}_{1,2}, \alpha(\overline{u}_{K_0} - \overline{u}_{K_1}), \alpha(\overline{u}_{K_3} - \overline{u}_{K_0})\right), \\ u_{3,4} &= \overline{u}_{K_0} + \mathcal{M}\left(\widetilde{u}_{2,3} - \overline{u}_{K_0}, \alpha(\overline{u}_{K_0} - \overline{u}_{K_1}), \alpha(\overline{u}_{K_3} - \overline{u}_{K_0})\right). \end{aligned}$$

Similarly, we reconstruct $u_{2,3}$ and $u_{4,1}$ in the y-direction by applying :

$$u_{2,3} = \overline{u}_{K_0} - \mathcal{M}\left(\overline{u}_{K_0} - \widetilde{u}_{2,3}, \alpha(\overline{u}_{K_0} - \overline{u}_{K_2}), \alpha(\overline{u}_{K_4} - \overline{u}_{K_0})\right),$$

$$u_{4,1} \quad = \quad \overline{u}_{K_0} + \mathcal{M}\left(\widetilde{u}_{4,1} - \overline{u}_{K_0}, \alpha(\overline{u}_{K_0} - \overline{u}_{K_2}), \alpha(\overline{u}_{K_4} - \overline{u}_{K_0})\right).$$

2. The aim now is to reconstruct the cell values at the vertices. Indeed, the midpoint edge averages $u_{i,j}$, which are already computed in the previous step, are not sufficient to uniquely reconstruct the nodal values u_i. Consequently, we use the following constraints which ensure the mass conservation over each edge within the cell :

$$\begin{aligned} u_1 + u_2 &= 2\,u_{1,2}, \\ u_2 + u_3 &= 2\,u_{2,3}, \\ u_3 + u_4 &= 2\,u_{3,4}, \\ u_4 + u_1 &= 2\,u_{4,1}. \end{aligned} \tag{10}$$

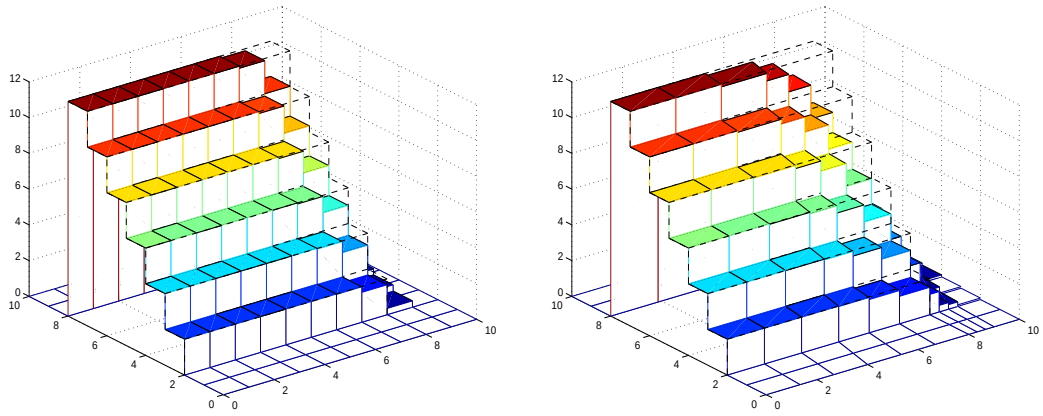
The values at the vertices u_i are thus reconstructed by combining the equality constrains (10) with those given in system (11). Therefore, the resulting slope limiter guarantees that no new extrema can be created at the vertices as well as at the midpoints of the edges within each cell. On the other hand, the obtained optimization problem is very simple to solve. The system of equality constraints (10) is of rank 3, thus the problem can be reduced to a minimization problem with only one variable. Further, the optimal solution can be attained without iterating.

4.1.1 Numerical test

The same test problem given in the previous section is now approximated by using the DG method with our modified slope limiter. Computations are done for a structured grid as well as for an unstructured grid of trapezoids. The time step is taken to be 0.62 for the unstructured mesh, so that the Courant number does not exceed 0.9 in the active elements of the grid. Results depicted in Fig.7 show that the modified slope limiter completely eliminates spurious oscillations with minimal numerical smearing, that is, with $\alpha = 1$.

4.2 Slope limiting for triangular elements

The construction of the slope limiting operator for triangular elements follows a similar approach used for rectangular elements. It is proved in [1] (see also [16]) that an appropriate local maximum principle is satisfied by ensuring that no new extrema are created at the midpoints of the grid edges. Consequently, the proposed slope limiting operator aims to eliminate oscillations at the midpoints of edges within each



$\alpha = 1$ $\alpha = 1$

Figure 7: DG results obtained when using the modified slop limiter for a structured and an unstructured grid.

cell. To describe the slope limiting procedure, let us consider a triangular element K_0 surrounded by its neighborhoods $K_i, i = 1, \cdots, 3$. (Fig. 8). The notations for the vertices and for the midpoints of the edges are the same as those used for quadrangles (Fig.6). The slope limiting process consists of two main operations :

1. The aim in the first stage is to reconstruct the average values $\widetilde{u}_{i,j}$ at the midpoints of the edges. An indispensable condition that must be satisfied is the local mass conservation. To obey a local maximum principle, some constraints are imposed to ensure that each reconstruction $u_{i,j}$ is between the cell averages of the two adjacent elements. To have a less restrictive limiting, the reconstructions $u_{i,j}$ are kept as close as possible to the initial state values $\widetilde{u}_{i,j}$. The resulting optimization problem to solve is therefore :

For given initial state values $\widetilde{U}_{\widehat{K_0}} = (\widetilde{u}_{1,2}, \widetilde{u}_{2,3}, \widetilde{u}_{3,1})$, find $U_{\widehat{K_0}}$ the solution of the problem :

$$\min_W \|W - \widetilde{U}_{\widehat{K_0}}\|_2, \quad \text{subject to the linear constraints :} \tag{11}$$
$$\overline{w} = \frac{1}{3}(w_{1,2} + w_{2,3} + w_{3,1}) = \overline{u}_{K_0},$$
$$(1-\alpha)\overline{u}_{K_0} + \alpha \min(\overline{u}_{K_1}, \overline{u}_{K_0}) \leq w_{1,2} \leq (1-\alpha)\overline{u}_{K_0} + \alpha \max(\overline{u}_{K_1}, \overline{u}_{K_0}),$$

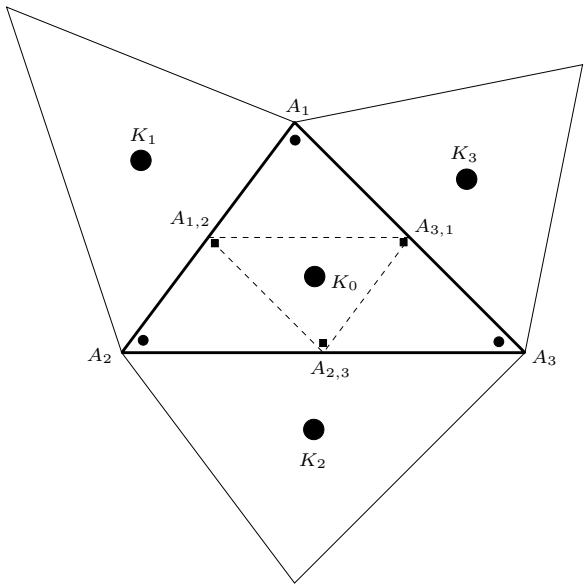


Figure 8: Illustration of limiting for triangular elements.

$$(1-\alpha)\overline{u}_{K_0} + \alpha \min(\overline{u}_{K_2}, \overline{u}_{K_0}) \leq w_{2,3} \leq (1-\alpha)\overline{u}_{K_0} + \alpha \max(\overline{u}_{K_2}, \overline{u}_{K_0}),$$
$$(1-\alpha)\overline{u}_{K_0} + \alpha \min(\overline{u}_{K_3}, \overline{u}_{K_0}) \leq w_{3,1} \leq (1-\alpha)\overline{u}_{K_0} + \alpha \max(\overline{u}_{K_3}, \overline{u}_{K_0}).$$

See appendix A for the solution of this problem.

2. Unlike the case of rectangular elements, the state values at the cell vertices can be directly computed by using the reconstructed midpoint edge values $u_{i,j}$. Therefore, a quite simple system of linear equation has to be solved :

$$\begin{aligned} u_1 + u_2 &= 2\,u_{1,2}, \\ u_2 + u_3 &= 2\,u_{2,3}, \\ u_3 + u_1 &= 2\,u_{3,1}. \end{aligned} \tag{12}$$

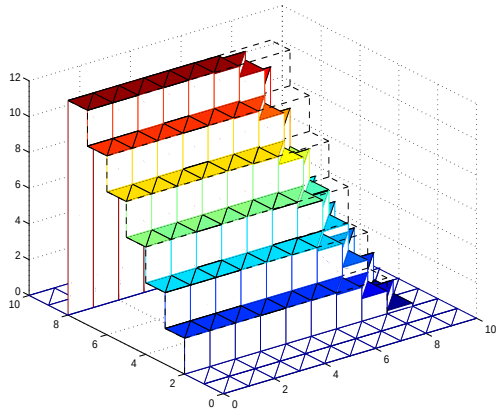
4.2.1 Degrees of freedom at the midpoints of edges

It seems evident that by defining the degrees of freedom at the midpoints of the grid edges the proposed slope limiter is more convenient since, in this case, no information at the vertices is required. Indeed, local approximation of the solution obtained by taking the degrees of freedom either at the vertices or at the midpoints of edges has

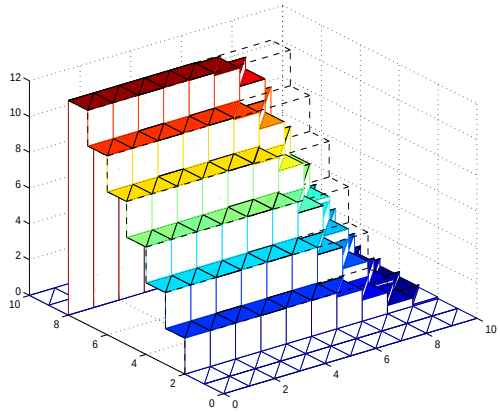
the same order of accuracy. However, we have found that the numerical solution obtained by the later approximation is more diffusive. This drawback is due to the upwinding. In fact by using the midpoints of edges as degrees of freedom the only information transmitted from one element to its neighbors, in the upwind direction, is the value at the midpoint of their common edge. On the other hand, by taking the degrees of freedom at the vertices, the two nodal values at the extremities of the common edges are transmitted. This approach is indeed more precise since it provides information about the state gradient along the edge.

4.2.2 Numerical test

In figure 9, we present the numerical results obtained by the DG method which is stabilized by using the modified slope limiter (limiter previously described). The obtained solution is free from any spurious oscillations even with minimal artificial diffusion ($\alpha = 1$). However, comparison between solutions obtained by the DG method with degrees of freedom at the vertices and at the midpoints of the edges, as illustrated in Fig.9, shows that the later approach is more diffusive.



Degrees of freedom at the grid vertices.



Degrees of freedom at the grid edge midpoints.

Figure 9: DG results for triangular grid with different degrees of freedom.

5 Existing slope limiters

In this section, we briefly review two slope limiters introduced by Cockburn and Shu in [9] for rectangular and triangular grids. We restrict the presentation for P^1 piecewise approximation functions.

5.1 Rectangular grids

The approximation solution $u_h(x, y, t)$ over each rectangular element $[x_{i-1/2}, x_{i+1/2}] \times [x_{i-1/2}, x_{i+1/2}]$ in a cartesian grid is approximated by means of P^1 polynomials. A convenient choice of the degrees of freedom is the cell average $\overline{u}$ and the two slopes of the state function u_x and u_y in the x- and y-directions, respectively. Thus, over each element we have :

$$u_h(x, y, t) = \overline{u}(t) + u_x(t)\phi_i(x) + u_y(t)\psi_i(t), \tag{13}$$

where

$$\phi_i(x) = \frac{x - x_i}{\Delta x/2}, \qquad \psi_i(x) = \frac{y - y_i}{\Delta y/2}.$$

5.1.1 Limiting

The reconstruction of u_x and u_y is carried out sequentially in the x- and y-directions by applying a one-dimensional slope limiter. Cockburn and Shu proposed to use the TVB *generalized* slope limiter [21, 7] for the reconstruction. Since our aim is to have the numerical solution free from any spurious oscillation, we will use the one-dimensional slope limiter proposed in this paper (6). Therefore, within each cell u_x and u_y are respectively replaced by :

$$\mathcal{M}\left(u_x, \alpha(\overline{u}_{i+1,j} - \overline{u}_{i,j}), \alpha(\overline{u}_{i,j} - \overline{u}_{i-1,j})\right),$$
$$\mathcal{M}\left(u_y, \alpha(\overline{u}_{i,j+1} - \overline{u}_{i,j}), \alpha(\overline{u}_{i,j} - \overline{u}_{i,j-1})\right).$$

5.2 Triangular grids

The approximation solution $u_h(x, y, t)$ is approximated by means of piecewise linear polynomials. The degrees of freedom are the state values at the midpoints of the grid edges.

5.3 Limiting

To describe the limiter, we use the same notations as in [9]. For an arbitrary triangle K_0 and its surrounding neighbors $K_i, i = 1, \cdots, 3$, the notations $b_i, i = 0, \cdots, 3$ and $m_i, i = 1, \cdots, 3$ refer respectively to the barycenters of the triangles and the midpoints of the edges within K_0 (see Fig. 10).

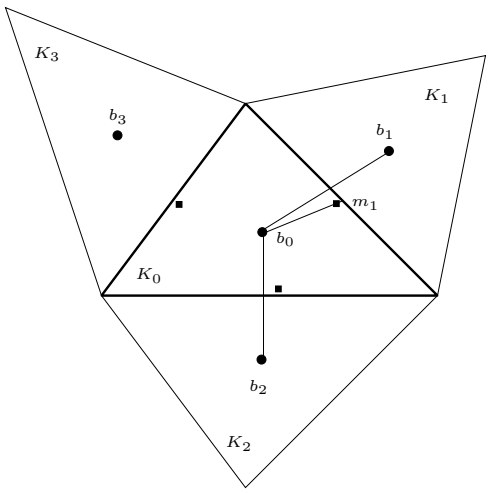


Figure 10: Cockburn and Shu limiting illustration for triangular elements.

Choosing any edge midpoint m_1, we get :

$$m_1 - b_0 = \alpha_1(b_1 - b_0) + \alpha_2(b_2 - b_0), \text{ for some } \alpha_1, \alpha_2 \in \mathbb{R}^2.$$

Then, for any *linear function* u_h we can write :

$$u_h(m_1) - u_h(b_0) = \alpha_1(u_h(b_1) - u_h(b_0)) + \alpha_2(u_h(b_2) - u_h(b_0)). \tag{14}$$

Since the cell average $\overline{u}_{K_i}$ is nothing but the value of the function at the barycenter $u_h(b_i)$, (14) is rewritten as follows :

$$\begin{aligned} \widetilde{u}_h(m_1, K_0) &\equiv u_h(m_1) - \overline{u}_{K_0} \\ &= \alpha_1(\overline{u}_{K_1} - \overline{u}_{K_0}) + \alpha_2(\overline{u}_{K_2} - \overline{u}_{K_0}) \\ &\equiv \Delta\overline{u}(m_1, K_0). \end{aligned} \tag{15}$$

To describe the slope limiting operator $\Lambda\Pi_h$, we consider any *piecewise linear function* u_h. Indeed equation (15) is no more valid. By using some basis functions ϕ_i, u_h can be expressed over K_0 as follows :

$$u_h(x,y) = \sum_{i=1}^{3} u_h(m_i)\phi_i(x,y) = \overline{u}_{K_0} + \sum_{i=1}^{3} \widetilde{u}_h(m_i,K_0)\phi_i(x,y).$$

First, we compute the quantities

$$\Delta_i = \mathcal{M}(\widetilde{u}_h(m_i,K_0), \nu\Delta\overline{u}(m_i,K_0)), \text{ for some } \nu > 1,$$

by using the *minmod* function described above. Note that Cockburn and Shu used instead their modified TVB *minmod* function. Consequently, reconstruction is carried out according to the following two cases :

1. If $\sum_{i=1}^{3} \Delta_i = 0$, we set

$$\Lambda\Pi_h u_h = \overline{u}_{K_0} + \sum_{i=1}^{3} \Delta_i \phi_i(x,y).$$

2. If $\sum_{i=1}^{3} \Delta_i \neq 0$, we compute

$$pos = \sum_{i=1}^{3} \max(0,\Delta_i), \qquad neg = \sum_{i=1}^{3} \max(0,-\Delta_i),$$

and define

$$\theta^+ = \min\left(1, \frac{neg}{pos}\right), \qquad \theta^- = \min\left(1, \frac{pos}{neg}\right).$$

Finally, we set

$$\Lambda\Pi_h u_h = \overline{u}_{K_0} + \sum_{i=1}^{3} \widehat{\Delta}_i \phi_i(x,y),$$

where

$$\widehat{\Delta}_i = \theta^+ \max(0,\Delta_i) - \theta^- \max(0,-\Delta_i).$$

This limiting operator conserves the mass within each element and guarantees that the reconstructed gradient of $\Lambda\Pi_h u_h$ is not bigger than that of u_h.

6 Numerical experiences

In order to test the behavior of the introduced slope limiters, we present two classical numerical experiments for linear convection equations with either rectangular of triangular grids. All the presented numerical tests are solved by using the simplified Runge-Kutta method for the time integration.

6.1 Diagonally moving prism

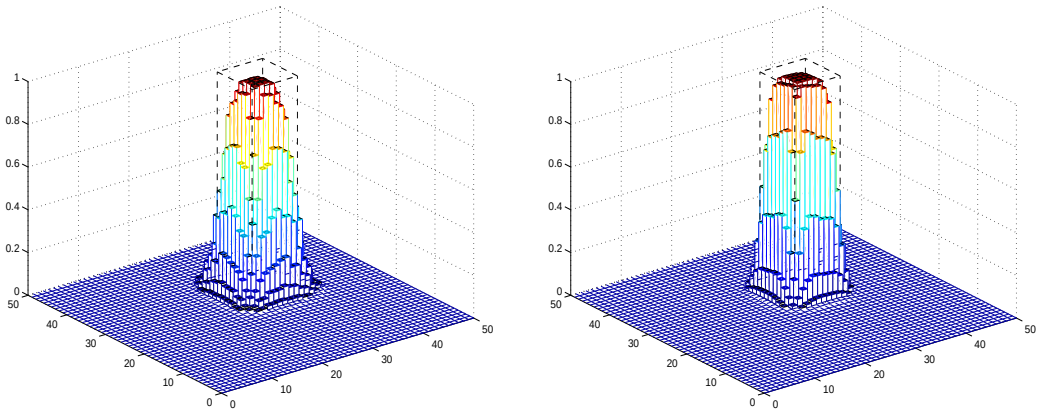
The first test is a solid shifting of a square along the diagonal of the computational domain $\Omega = (0, 50) \times (0, 50)$. The initial profile is given by :

$$u_h(x, y, 0) = \begin{cases} 1 & (x, y) \in \quad [1, 10] \times [1, 10], \\ 0 & elsewhere. \end{cases}$$

A grid of 50×50 rectangular elements is considered to test the different reconstruction techniques. Periodic boundary conditions are applied. A parallel flow is taken diagonal to the grid such that $\beta = (1/2, 1/2)$. The time step is chosen to fix the condition $\mathcal{C} = \beta_x \Delta t/\Delta x = \beta y \Delta t/\Delta_y$. In figures 11, 12 and 13, we present the DG results obtained by using Chavent-Jaffré, Cockburn-Shu and the modified slope limiters, respectively. The dashed lines represents the shifted profile after a simulation time T=50. It is clear that the slope limiter introduced by Chavent and Jaffré (Fig.11) produces some oscillations. However decreasing the time step leads to some smoothing in the solution. One the other hand, results obtained by Cockburn-Shu and the modified slope limiter seam to be very comparative. Table 1 presents the L_1 and L_2 errors against the resolution for the DG numerical solutions by using the three slope limiters. The modified slope limiter gives slightly better accuracy than the others.

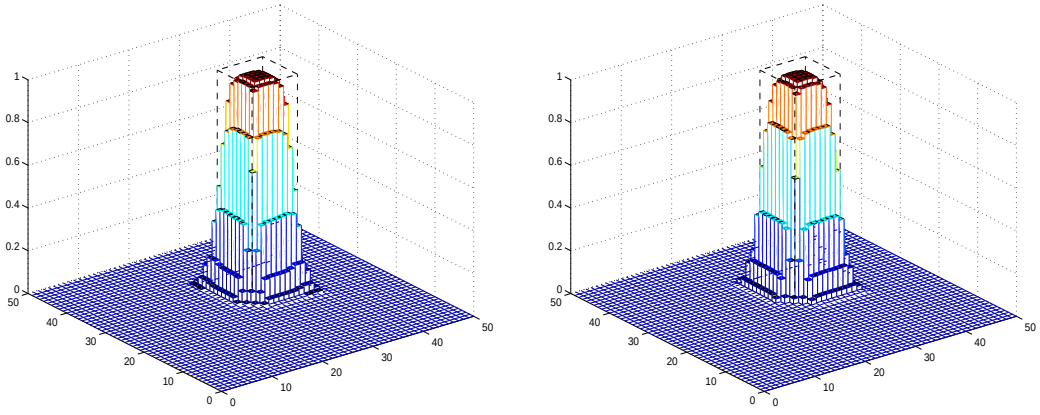
Table 1: L_1 and L_2 relative errors for different slope limiters.

Slope limiter	$\mathcal{C} = 0.6$		$\mathcal{C} = 0.1$	
	$10^2.error$		$10^2.error$	
	$L_1 - error$	$L_\infty - error$	$L_1 - error$	$L_\infty - error$
Chavent-Jaffré	3.41	37.95	2.80	27.46
Cockburn-Shu	2.82	28.03	2.67	26.67
Modified-limiter	2.72	27.27	2.50	23.34



$\mathcal{C} = 0.6$ $\mathcal{C} = 0.1$

Figure 11: Results obtained by using the slope limiter introduced by Chavent and Jaffré with two different time steps.

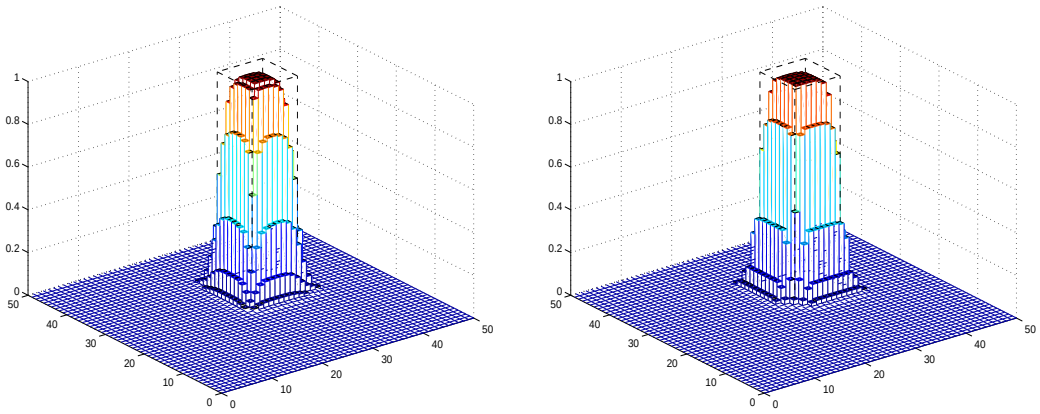


$\mathcal{C} = 0.6$ $\mathcal{C} = 0.1$

Figure 12: Results obtained by using the slope limiter introduced by Cockburn and Shu with two different time steps.

6.2 Rotating cylinder

A classical test for multi-dimensional scalar convection equation is the rotating cylinder (see, e.g. [18]). Thus, a circular peak is moved in a rotating flow field. The



$\mathcal{C} = 0.6$ $\mathcal{C} = 0.1$

Figure 13: Results obtained by using the modified slope limiter with two different time steps.

results after four rotations are compared with the exact solution which is simply the initial condition. Three grids made of rectangles, parallelograms and arbitrary triangles are used to examine the behavior of the described slope limiters. The rotational velocity field $\beta(x) = r(x)(-\sin\theta, \cos\theta)$ is one rotation in $T = 2\pi$, where $r(x) = \|x - x_0\|$ is the rotation radius around the center x_0. The time discretization step is taken to be 0.01, so that the Courant number does not exceed unity everywhere in the domain.

In the first test, the computational domain $\Omega = (0, 50) \times (0, 50)$ is discretized into a cartesian grid of 100×100 cells. Figures 14, 15 and 16 display the profile and the isolines of the DG solutions obtained by using Chavent-Jaffré, Cockburn-Shu and the modified slope limiters. The first limiter gives the least accurate solution. On the other hand, solutions obtained by the two other limiters seem to be very similar. Nevertheless, the L_1 and L_2 errors given in Fig.17 show a slight accuracy of the modified limiter.

In the second test, the domain is discretized into a grid of parallelograms. In figure 18, we present the results obtained after four rotations of the initial profile obtained by using Chavant-Jaffré and the modified slope limiters. The first limiter clearly suffers from dispersive errors.

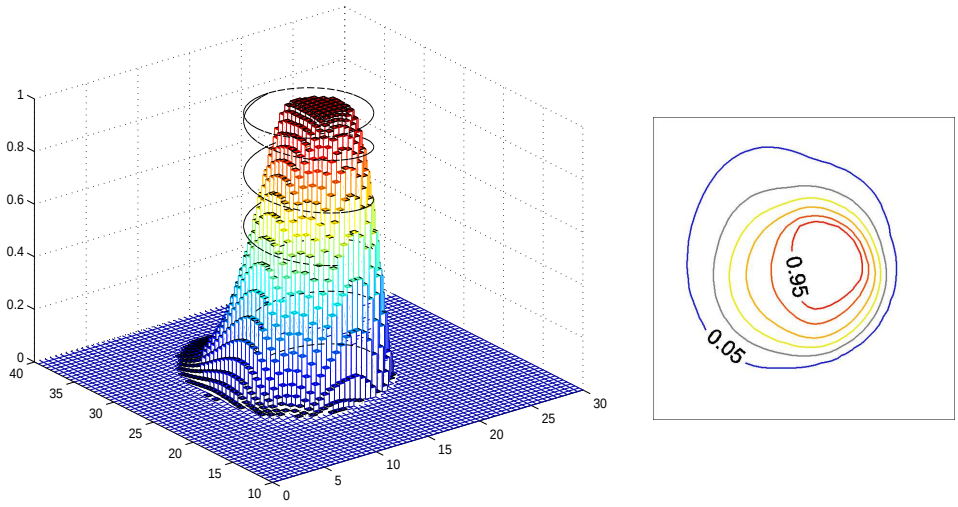


Figure 14: Profile and isolines of the DG solution obtained by using Chavent-Jaffré slope limiter after four rotations.

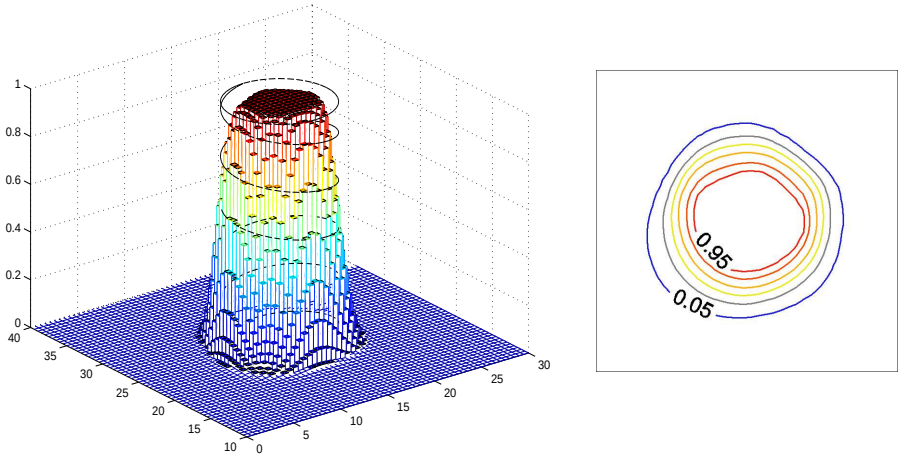


Figure 15: Profile and isolines of the DG solution obtained by using Cockburn-Shu slope limiter after four rotations.

In the final test, computations are carried out on an arbitrary grid of triangular elements made of 8000 cells. The results obtained by Chavent-Jaffré, Cockburn-Shu

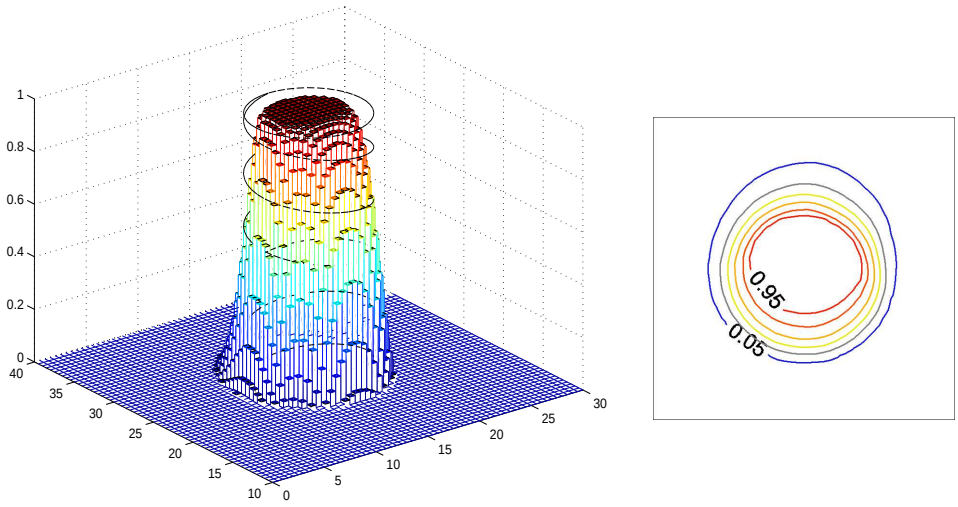


Figure 16: Profile and isolines of the DG solution obtained by using the modified slope limiter after four rotations.

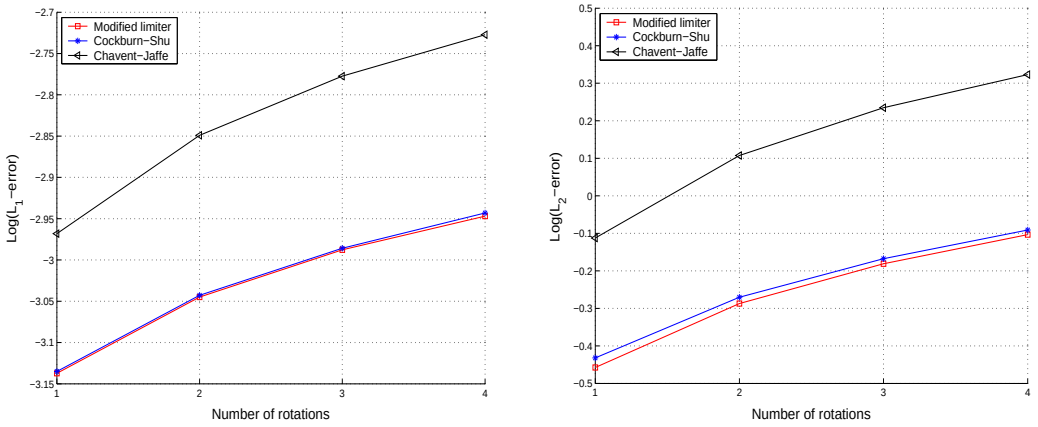
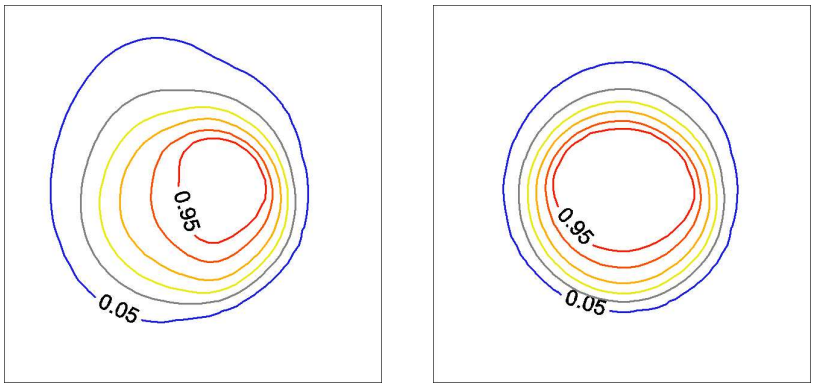


Figure 17: L_1 and L_2 errors for the rotating cylinder.

and the modified slope limiters are depicted in figures 19, 20 and 21, respectively. Errors presented in Fig.23 show that the limiter for triangular grids introduced by Cockburn and Shu is the least accurate. It should be noted that the degrees of freedom are chosen at the grid vertices. As we have previously mentioned, the DG



Chavent-Jaffré limiter Modified limiter

Figure 18: Results after four rotations obtained by using Chavent-Jaffré and the modified slope limiters on a grid made of parallelograms.

method generates excessive smearing when degrees of freedom of the state function are sought at the midpoints of the grid edges. In Fig.22, we present the DG approximation solution obtained when using the modified slope limiter. Further, the two other slope limiters produce similar results.

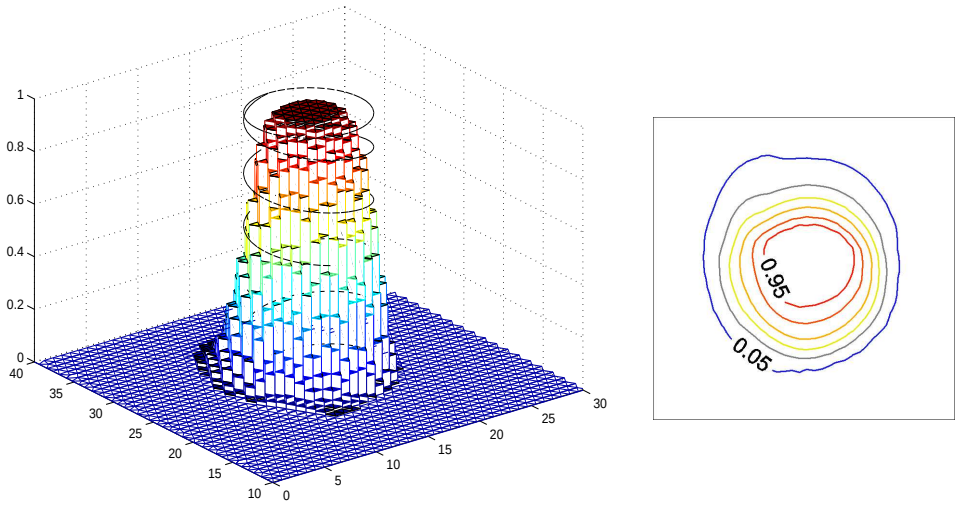


Figure 19: Profile and isolines of the solution obtained by using Chavent-Jaffré slope limiter over a triangular grid.

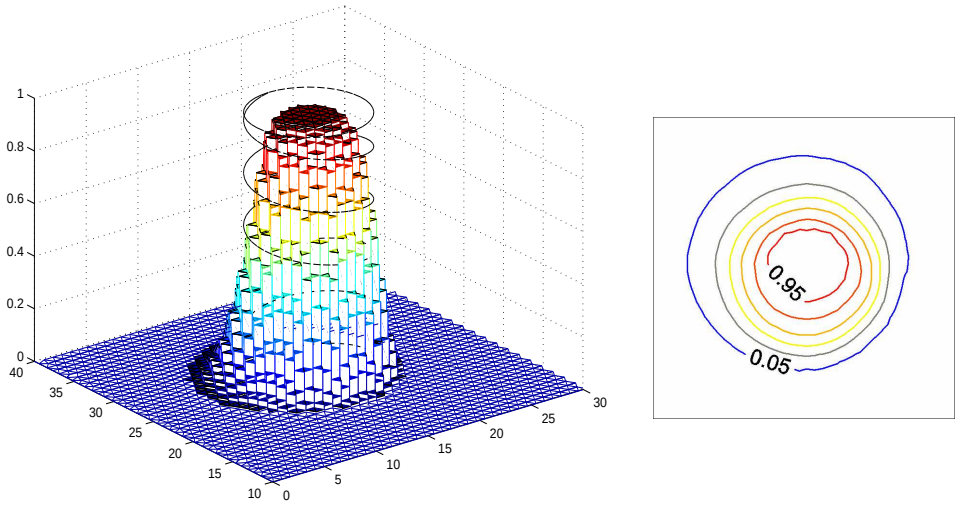


Figure 20: Profile and isolines of the solution obtained by using Cockburn-Shu slope limiter over a triangular grid.

7 Conclusion

Data reconstruction is crucial for the stabilization of high order discontinuous Galerkin methods. In one-dimensional space, many successful slope limiters have been de-

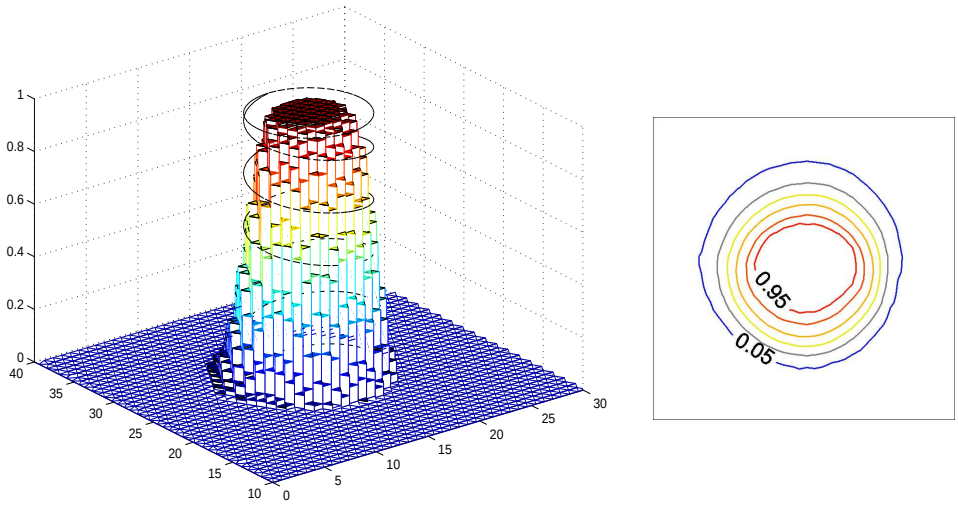


Figure 21: Profile and isolines of the solution obtained by the modified limiter over a triangular grid.

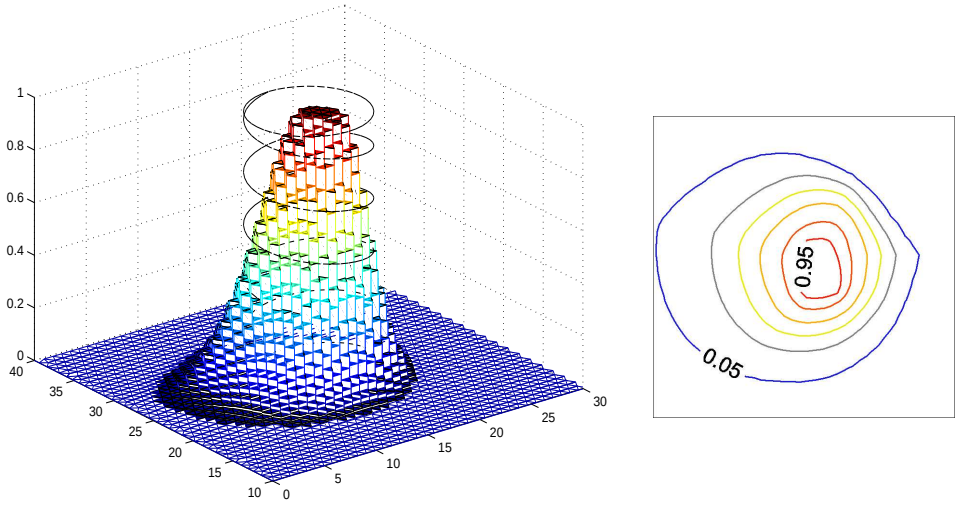


Figure 22: The approximation solution obtained by the DG method with degrees of freedom at the midpoints of the grid edges.

veloped. However, in higher dimensions, specially on unstructured grids, the construction of reliable slope limiters that preserve the accuracy of the scheme is still a

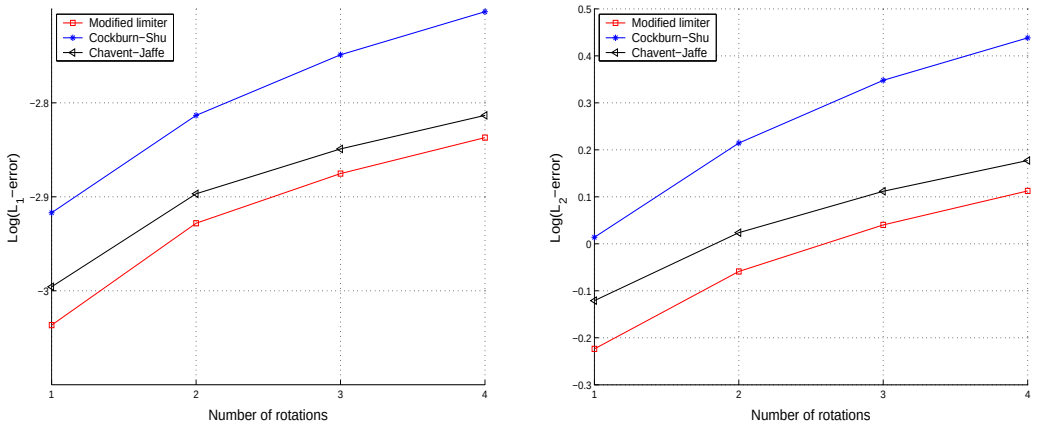


Figure 23: L_1 and L_2 errors for the rotating cylinder on the triangular mesh.

challenge.
The multi-dimensional slope limiter introduced by Chavent and Jaffré constitutes the bulk of this work. This limiter is considered as an extension of the Van Leer's MUSCL slope limiter. We have given some numerical tests for both rectangular and triangular discretizations where this limiter fails to eliminate all spurious oscillations. This drawback is due to the fact that the reconstruction of data by means of local constraints at the vertices is insufficient to prevent unphysical values at the midpoints of the edges. The proposed remedy is to reconstruct data by using constraints applied at the midpoints of the grid edges. This approach has an important physical property since it limits the numerical fluxes across the interelements rather than the function values at the grid vertices.
For rectangular elements, piecewise quadratic functions are used for the approximation space. The solution is reconstructed first at the midpoints of the cell edges by means of a dimension splitting technique. The nodal values are then reconstructed by solving a minimization problem. Similar approach is used to limit slopes for triangular grids. However, we have found that by taking the degrees of freedom of the approximation solution at the midpoints of the edges, the scheme becomes more diffusive.
Numerical comparisons with other slope limiters showed a good improvement of the proposed reconstruction techniques.

Appendix A

The minimization problem described in section 3.2 may introduce some difficulties for the resolution. This problem is rewritten as follows :

For a given vector $\widetilde{U}_K \in \mathcal{P}_K$, find $U_K \in \mathcal{P}_K \cap \mathcal{Q}_K$ the solution of the least squares problem :

$$\min_{W \in \mathcal{P}_K \cap \mathcal{Q}_K} \mathcal{J}(W),$$

where $\mathcal{J}(W) = \frac{1}{2}\|W - \widetilde{U}_K\|_2$ is the objective function, $\mathcal{P}_K$ and $\mathcal{Q}_K$ are the hyperplane and the hypercube describing respectively the linear equality and inequality constraints as following :

$$\begin{aligned} \mathcal{P}_K &= \left\{ W \in \mathbb{R}^{\mathcal{N}_K}; \sum_{i=1}^{\mathcal{N}_K} w_i = \mathcal{N}_K \overline{w}_K \right\}, \\ \mathcal{Q}_K &= \prod_{i=1}^{\mathcal{N}_K} [\gamma_i \, , \, \mu_i] \, , \end{aligned}$$

with $\gamma_i = (1-\alpha)\overline{u}_K + \alpha \overline{u}_{\min,i}, \quad \mu_i = (1-\alpha)\overline{u}_K + \alpha \overline{u}_{\max,i}$.

It is easy to check that the convex closed set $\mathcal{P}_K \cap \mathcal{Q}_K$ is non empty since it contains the point $W = (w_i = \overline{u}_k, i = 1, \cdots, \mathcal{N}_K)$. Thus, the convex property of the objective function guarantees the existence and the uniqueness of the solution.

In this appendix, we present an efficient algorithm which is based on the so-called *active set algorithm* [3]. This algorithm is not iterative in nature, but rather it decreases the value of the objective function so that the optimal solution is attained in finitely many steps.

For any $W \in \mathbb{R}^{\mathcal{N}_K}$, W is a *feasible point* if it satisfies all the inequality constraints, i.e., $W \in \mathcal{Q}_K$. Let us denote by $I = I(W)$ the set of indices of active constraints at W, i.e., constraints satisfied with equality.

In the active set algorithm a sequence of equality-constrained problems are solved corresponding to a prediction of the active set. At each step one constraint is added to or dropped from the active set. The stepping process is described as follows :

1. Initialization :
Choose $W^{(0)} = (\overline{w}_K, i = 1, \cdots, \mathcal{N}_K)$ a feasible point.

2. Stepping process :
Let $W^{(k)}$ be the iterate at the k-th step and $I^{(k)}$ the corresponding active set.
The process seeks $(W^{(k+1)}, I^{(k+1)})$ according to the following steps :

3. Solve the equality-constrained problem :

$$\begin{cases} \min\limits_{Z} \mathcal{J}(Z) \quad \text{subject to} \\ Z \in \mathcal{P}_K, \\ z_i = w_i^{(k)} \text{ for } i \in I^{(k)}. \end{cases}$$

This problem can be easily solved by using Lagrange multipliers λ_i.

4. If Z is a feasible point, then

Check for optimality such that :

$$\forall i \in I^{(k)}, \lambda_i \text{ is optimal} \Leftrightarrow \begin{cases} (w_i^{(k)} = \gamma_i \quad \text{and} \quad \lambda_i \leq 0), \\ \text{or} \\ (w_i^{(k)} = \mu_i \quad \text{and} \quad \lambda_i \geq 0). \end{cases}$$

4.1. If $I^{(k)} = \emptyset$ or $\lambda_i, i \in I^{(k)}$, are optimal then
the optimal solution Z is reached.

4.2. Otherwise, there exists $i \in I^{(k)}$ such that λ_i is non optimal, then
one active constraint is dropped such that
$I^{(k+1)} = I^{(k)} \setminus \{i\}$, where $|\lambda_i| = \max\{|\lambda_j|; \lambda_j\}$ is non optimal.
Set $W^{(k+1)} = Z$.
Go to step 2.

5. If Z is not feasible then
Choose $\delta = \min\left\{\delta_i; i \notin I^{(k)}\right\} \in [0, 1[$, such that

$$\delta_i = \begin{cases} \dfrac{w_i^{(k)} - \gamma_i}{w_i^{(k)} - z_i}, & \text{if} \quad z_i < \gamma_i \\ \\ \dfrac{w_i^{(k)} - \mu_i}{w_i^{(k)} - z_i}, & \text{if} \quad z_i < \mu_i. \end{cases}$$

Set $W^{(k+1)} = W^{(k)} + \delta\left(Z - W^{(k)}\right)$.
Update $I^{(k+1)}$ by checking the active constraints.
Go to step 2.

This algorithm is not expensive from a computational point of view. Numerical observations showed that the optimal solution is reached with at most $2\mathcal{N}_K$ steps.

References

[1] P. Batten, C. Lambert and D. Causen. Positively conservative high-resolution convection schemes for unstructured elements, *Int. J. Numer. Methods Eng.*, 39:1821–1838, (1996).

[2] M. Buès and C. Oltean. Numuerical simulations for saltwater intrusion by mixed hybrid finite element method and discontinuous finite element method, *Transport in Porous Media*, 40:171–200, (2000).

[3] A. Bjorck. Numerical methodes for least squares problems, *SIAM*, Philadelphia, (1996).

[4] G. Chavent and J. Jaffré. Mathematical Models and Finite Elements for Reservoir Simulation, *Studies in Mathematics and its applications*, North Holland, Amsterdam, (1986).

[5] G. Chavent and B. Cockburn. The local projection P^0 P^1 -discontinuous Galerkin finite element method for scalar conservation laws, *M2AN*, 23:565–592, (1989).

[6] G. Chavent and Salzano. A finite-element method for the 1-D water flooding problem with gravity, *J. Comput. Phys.*, 45:307–344, (1982).

[7] B. Cockburn and C.W. Shu. TVB Runge Kutta local projection discontinuous Galerkin finite element method for conservative laws II: General frame-work, *Math. Comp.*, 52:411–435, (1989).

[8] B. Cockburn, S. Hou and C.W. Shu. TVB Runge Kutta local projection discontinuous Galerkin finite element method for conservative laws III: One dimensional systems, *J. Comput. Phys.*, 84:90–113, (1989).

[9] B. Cockburn and C.W. Shu. The Runge-Kutta Discontinuous Galerkin Method for conservative laws V: Multidimentional Systems, *J. Comput. Phys.*, 141:199–224, (1998).

[10] V. Gowda. Discontinuous finite elements for nonlinear scalar conservation laws, *Thèse de Doctorat*, Université Paris IX, (1988).

[11] V. Gowda and J. Jaffré. A discontinuous finite element method for scalar nonlinear conservation laws, *Rapport de recherche INRIA*, N^o 1848, (1993).

[12] S. Godunov. Finite difference methods for numerical computation of discontinuous solutions of the equations of fluid dynamics, *Math. Sbornik* , 47:271–306 (1959).

[13] J. Goodman and R. LeVeque. On the accuracy of stable schemes for 2D conservation laws, *Math. Comp.*, 45:15–21, (1985).

[14] A. Harten. High resolution schemes for hyperbolic conservation laws, *J. Comput. Phys.*, 49:357–393, (1983).

[15] A. Harten. On a class of high resolution total-variation-stable finite-difference schemes, *SIAM J. Numer. Anal.*, 21:1–23 (1984).

[16] M.E. Hubbard. Multidimensional Slope Limiters for MUSCL-Type Finite Volume Schemes on Unstructured Grids, *J. Comput. Phys.*, 155:54–74, (1999).

[17] C. Hirsch. Numerical Computation of Internal and External Flows, *A Wiley-Interscience Publication*, New York, (1990).

[18] C. Johnson. Numerical solution of partial differential equations by the finite element method, *Cambridge university press*, (1995).

[19] L. Kaddouri. Une méthode d'élément finis discontinus pour les équations d'Euler des fluides compressibles, *Thèse de Doctorat*, Université Paris VI, (1988).

[20] S. Osher. Convergence of generalized MUSCL schemes, *SIAM J. Numer. Anal.*, 28:907–922, (1993).

[21] C.W. Shu. TVB uniformly high order schemes for conservative laws, *Math. Comp.*, 49:105–121, (1987).

[22] P. Siegel , R. Mosé, Ph. Ackerer and J. Jaffre. Solution of the advection dispersion equation using a combination of discontinuous and mixed finite elements, *Journal for Numerical Methods in Fluids*, 24:595–613, (1997).

[23] E. Toro, Riemann Solvers and Numerical Methods for Fluid Dynamics, *Springer*, Berlin, (1997).

[24] B. Van Leer. Towards the ultimate conservative scheme: II., *J. Comput. Phys.*, 14: 361–376, (1974).

[25] B. Van Leer. Towards the ultimate conservative scheme: IV. A new approch to numerical convection, *J. Comput. Phys.*, 23:276–299, (1977).

[26] B. Van Leer. Towards the ultimate conservative scheme: V. A second order Godunov's method, *J. Comput. Phys.*, 32:101–136, (1979).

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399