Limiting strategies for polynomial reconstructions in Finite Volume approximations of the linear advection equation

Enrico Bertolazzi[1] **and Gianmarco Manzini**[2]
[1]Dip. Ingegneria Meccanica e Strutturale
Università di Trento, via Mesiano 77, 38050 Trento, Italy
enrico.bertolazzi@ing.unitn.it
[2]Istituto di Analisi Numerica IAN – CNR
via Ferrata 1, I – 27100 Pavia, Italy
marco.manzini@ian.pv.cnr.it

Abstract

A second-order accurate cell-centered Finite Volume method is proposed to solve the time-dependent scalar advection equation. The spatial accuracy is ensured by a piecewise linear reconstruction which requires a suitable limiting strategy to control spurious numerical oscillations. Three different approaches are analysed to limit the approximate solution.

Keywords: Finite Volumes, non-oscillatory reconstructions, limiters

1. Introduction

The solution to the scalar advection equation is approximated in the framework on the Finite Volume (FV) methods. An FV scheme is proposed, which formally attains second-order accuracy by a piecewise-linear reconstruction from cell-averages. A limiter is introduced to ensure a non-linear stability condition. We propose and analyse three different limiting constraints.

The outline of the paper follows. The model problem is described in section 2, while the Finite Volume formulation and the piecewise-linear reconstruction algorithm are detailed in section 3. The first limiting condition that we consider in this work is presented and discussed in section 4. The method is reformulated in a more compact matrix-like form in section 5, while two stronger limiting constraints are introduced and their consequences analised in section 6. Finally, conclusions are reported in section 7.

2. The Scalar Advection Equation

Our model problem is the scalar advection equation, which reads as

$$C_t + \nabla \cdot (C\mathbf{V}) = 0 \qquad \text{on} \quad \Omega \tag{1}$$

In equation (1) $C(t, \mathbf{x})$ is a scalar quantity advected by an assigned constant velocity field $\mathbf{V}$ throughout a closed and connected domain Ω. The boundary of Ω, indicated by $\partial\Omega$, can be split into an inflow and an outflow part defined by

$$\begin{aligned} \text{inflow boundary}: \qquad & \Gamma^- = \{\mathbf{x} \in \partial\Omega \mid v_n < 0\}, \\ \text{outflow boundary}: \qquad & \Gamma^+ = \{\mathbf{x} \in \partial\Omega \mid v_n \geq 0\}, \end{aligned}$$

where $v_n = \mathbf{V} \cdot \mathbf{n}$, and such that $\partial\Omega = \Gamma^- \cup \Gamma^+$. The model problem described by equation (1) is completed by a suitable *initial solution*

$$C(0, \mathbf{x}) = C_0(\mathbf{x}), \qquad \mathbf{x} \in \Omega,$$

and a *Dirichlet boundary condition* on the inflow boundary, that is

$$C(t, \mathbf{x}) = \mathrm{g}(t, \mathbf{x}), \qquad \text{for} \quad (t, \mathbf{x}) \in \mathbb{R}^+ \times \Gamma^-(t).$$

We assume that

$$\mathrm{g}(t, \mathbf{x}) \geq 0, \quad \text{for } (t, \mathbf{x}) \in \mathbb{R}^+ \times \partial\Omega, \tag{2}$$

and

$$C_0(\mathbf{x}) \geq 0, \quad \text{for } \mathbf{x} \in \Omega. \tag{3}$$

Under assumptions (2-3), the following properties hold [1]:

$$\begin{aligned} &(a) \quad 0 \leq C(t, \mathbf{x}) \leq M(t), \quad \mathbf{x} \in \mathbb{R}^+ \times \Omega \\ &(b) \quad \frac{d}{dt} \| C \|_{L^1(\Omega)} + \langle \gamma C, v_n \rangle = 0 \\ &(c) \quad \frac{d}{dt} \| C \|^2_{L^2(\Omega)} + \left\langle (\gamma C)^2, v_n \right\rangle = 0 \end{aligned} \tag{4}$$

where

$$M(t) = \max \left\{ \| C_0 \|_{L^\infty(\Omega)}, \sup_{\tau \in [0,t]} \| \mathrm{g}(\tau, \cdot) \|_{L^\infty(\Gamma^-(\tau))} \right\}$$

and the symbol $\langle u, v \rangle$ denotes the usual scalar product for the two scalar functions u and v, i.e.

$$\langle u, v \rangle = \int_{\partial\Omega} u(\cdot, \mathbf{x})\, v(\cdot, \mathbf{x})\, d\mathsf{s}.$$

3. The Finite-Volume Formulation

Let us introduce a triangulation $\mathcal{T}_h$ of the domain Ω, which is supposed, as usual, regular and conformal in the sense specified by [2]. The Finite-Volume approximation of equation (1) logically proceeds in two steps. In the first step, equation (1) is re-formulated on the triangulated domain by integrating within each macroscopic control volume – also called *cell* – and then by applying the Gauss divergence theorem [3]. We have

$$\frac{\partial}{\partial t}\int_{\mathsf{T}_i} C(\cdot,\mathbf{x})\,d\mathbf{x} + \sum_{j\in\mathcal{T}_h(i)\cup\mathcal{T}_h'(i)} \int_{\mathsf{e}_{ij}} C(\cdot,\mathbf{x})\mathbf{V}\cdot\mathbf{n}_{ij}\,d\mathsf{s} = 0, \qquad \text{for every} \quad \mathsf{T}_i \in \mathcal{T}_h, \tag{5}$$

where we also introduced the following symbols:

- e_{ij} is the edge shared by the two control volumes T_i and T_j, i.e. $\mathsf{e}_{ij} = \mathsf{T}_i \cap \mathsf{T}_j$;
- $\mathcal{T}_h(i)$ is the set of volumes adjacent to the cell T_i; that is, for any $j \in \mathcal{T}_h(i)$ there exists a mesh cell T_j sharing the edge e_{ij} with T_i;
- $\mathcal{T}_h'(i)$ is the set of boundary edges of the cell T_i; that is, $\mathsf{e}_{ij'} = \mathsf{T}_i \cap \partial\Omega$ is an edge of the triangulation for any $j' \in \mathcal{T}_h'(i)$.

The second step consists in approximating (5) by the following relation

$$|\mathsf{T}_i|\,\frac{dc_i}{dt} + \sum_{j\in\mathcal{T}_h(i)} \mathrm{G}_{ij}^h(\mathcal{R}(\cdot,\cdot;\mathbf{c})) + \sum_{j\in\mathcal{T}_h'(i)} \mathrm{F}_{ij}^h(\mathcal{R}(\cdot,\cdot;\mathbf{c})) = 0, \tag{6}$$

which holds for every $t > 0$ and for every $\mathsf{T}_i \in \mathcal{T}_h$. In equation (6), c_i approximates the cell-averaged value of the advected quantity C,

$$c_i \approx \frac{1}{|\mathsf{T}_i|}\int_{\mathsf{T}_i} C(\cdot,\mathbf{x})\,d\mathbf{x},$$

while G_{ij}^h and F_{ij}^h are the approximate integrals of the numerical flux functions defined respectively for the internal and boundary edges,

$$\mathrm{G}_{ij}^h \approx \int_{\mathsf{e}_{ij}} C(\cdot,\mathbf{x})\mathbf{V}\cdot\mathbf{n}_{ij}\,d\mathsf{s}, \qquad j \in \mathcal{T}_h(i),$$

$$\mathrm{F}_{ij}^h \approx \int_{\mathsf{e}_{ij}} \mathrm{g}(\cdot,\mathbf{x})\mathbf{V}\cdot\mathbf{n}_{ij}\,d\mathsf{s}, \qquad j' \in \mathcal{T}_h^D(i).$$

Both these integrals are computed by the midpoint quadrature rule and the first one requires the cell-interface values of the FV approximate solution. A piecewise linear approximation of the solution within each cell is computed by using a suitable reconstruction procedure [5], thus obtaining

$$\mathcal{R}_i(\cdot,\mathbf{x};\mathbf{c}) = c_i + \mathcal{G}_i(\mathbf{c})\cdot(\mathbf{x}-\mathbf{x}_i), \qquad \mathbf{x}\in\mathsf{T}_i. \tag{7}$$

In equation (7) the term $\mathcal{G}_i(\mathbf{c})$ is the cell-centered reconstructed gradient, which is defined consistently to the solution cell-averages,

$$c_i = \frac{1}{|\mathsf{T}_i|}\int_{\mathsf{T}_i} \mathcal{R}_i(\cdot\,,\mathbf{x};\mathbf{c})\,d\mathbf{x}, \qquad \text{for every } \mathsf{T}_i \in \mathcal{T}_h.$$

Let us now introduce the edge-integrated velocity

$$\nu_{ij} = \int_{\mathsf{e}_{ij}} \mathbf{V}\cdot \mathbf{n}_{ij}\, d\mathsf{s}, \qquad \text{for every } \mathsf{e}_{ij} \in \mathcal{E}_h.$$

Applying a standard upwind technique [4], we have

$$\begin{aligned} \mathrm{G}_{ij}^h(\mathbf{c}) &= \mathcal{R}_i(\cdot\,, \mathbf{x}_{ij}; \mathbf{c})\nu_{ij}^+ + \mathcal{R}_j(\cdot\,, \mathbf{x}_{ij}; \mathbf{c})\nu_{ij}^-, \\ &= \mathcal{R}_i(\cdot\,, \mathbf{x}_{ij}; \mathbf{c})\nu_{ij}^+ - \mathcal{R}_j(\cdot\,, \mathbf{x}_{ij}; \mathbf{c})\nu_{ji}^+, \qquad j \in \mathcal{T}_h(i), \end{aligned} \tag{8}$$

where $\nu_{ij}^{\pm} = (\nu_{ij}^+ \pm \left|\nu_{ij}^+\right|)/2$ and we also used the fact that $\nu_{ij}^+ = -\nu_{ji}^+$. The integral of the numerical flux at the boundary edge e_{ij} is given by

$$\mathrm{F}_{ij}^h(\mathbf{c}) = \nu_{ij} \begin{cases} \mathrm{g}(\cdot, \mathbf{x}_{ij}), & \text{if } \nu_{ij} < 0; \\ \mathcal{R}_i(\cdot, \mathbf{x}_{ij}, \mathbf{c}), & \text{otherwise.} \end{cases}$$

4. A First Condition on the Limiter

In order to control the numerical oscillations, we assume that a limiter is introduced into the reconstruction procedure [5]. The limited reconstructed solution within T_i is again denoted by $\mathcal{R}_i$ and is obtained by the following process:

$$\mathcal{R}_i(\cdot, \mathbf{x}; \mathbf{c}) \longleftarrow c_i + \ell_i(\mathcal{R}_i(\cdot, \mathbf{x}; \mathbf{c}) - c_i). \tag{9}$$

The scalar factor ℓ_i in (9) is defined as the maximum value in $[0, 1]$ such that for each adjacent edge e_{ij} there exists a non-negative scalar coefficient $\sigma_{ij}(\mathbf{c})$,

$$0 \le \sigma_{ij}(\mathbf{c}) < \infty. \tag{COND.1}$$

The limited cell-interface reconstructed solution $\mathcal{R}_i(\cdot\,, \mathbf{x}_{ij}; \mathbf{c})$ can thus be reformulated as:

$$\mathcal{R}_i(\cdot, \mathbf{x}; \mathbf{c}) = c_i + \sigma_{ij}(\mathbf{c}) \times \begin{cases} c_j - c_i, & j \in \mathcal{T}_h(i) \\ \mathrm{g}_{ij} - c_i & j \in \mathcal{T}_h'(i) \end{cases}$$

5. The Semi-Discrete Formulation

Assuming condition (COND.1) and introducing the scalar quantity

$$\widetilde{\nu}_{ij}(\mathbf{c}) = \nu_{ij}^+\sigma_{ij}(\mathbf{c}) + \nu_{ji}^+\sigma_{ji}(\mathbf{c}),$$

in (8) we have

$$\mathrm{G}_{ij}^h(\mathbf{c}) = \nu_{ij}^+ c_i - \nu_{ji}^+ c_j + \widetilde{\nu}_{ij}(\mathbf{c})(c_j - c_i), \qquad j \in \mathcal{T}_h(i),$$

so that the balance of the edge fluxes for the i-th cell in (6) can be written as

$$\sum_{j\in\mathcal{T}_h(i)} \mathrm{G}_{ij}^h(\mathbf{c}) = \Big(\mathbf{G}\mathbf{c} - \widetilde{\mathbf{G}}(\mathbf{c})\mathbf{c}\Big)\Big|_i, \qquad \text{for each } \mathsf{T}_i \in \mathcal{T}_h,$$

where

$$\mathbf{G}|_{ij} = \begin{cases} -\nu_{ji}^{+} & \text{if } j \in \mathcal{T}_h(i), \\ \sum\limits_{k \in \mathcal{T}_h(i)} \nu_{ik}^{+} & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases} \qquad \widetilde{\mathbf{G}}|_{ij}(\mathbf{c}) = \begin{cases} -\widetilde{\nu}_{ji}(\mathbf{c}) & \text{if } j \in \mathcal{T}_h(i), \\ \sum\limits_{k \in \mathcal{T}_h(i)} \widetilde{\nu}_{ik}(\mathbf{c}) & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases}$$

the contribution to the balance of the boundary fluxes is indicated by

$$\sum_{j \in \mathcal{T}_h'(i)} \mathrm{F}_{ij}(\mathbf{c}) = \mathbf{f}(\mathbf{c})|_i, \qquad \text{for each } \mathsf{T}_i \in \mathcal{T}_h.$$

Assuming (COND.1), the FV semi-discrete formulation takes the form

$$\mathbf{T}\frac{d\mathbf{c}}{dt} + \mathbf{f}(\mathbf{c}) + \mathbf{G}\mathbf{c} - \widetilde{\mathbf{G}}(\mathbf{c})\mathbf{c} = \mathbf{0},$$

where $\mathbf{T}$ is the mass matrix and the term $\mathbf{f}(\mathbf{c})$ contains the contribution of the boundary conditions. It is possible to show that [1]

- $\mathbf{G}$ is a singular M-matrix;
- $\widetilde{\mathbf{G}}(\mathbf{c})$ is a singular Stieltjes matrix.

6. Stronger Constraints in the Limiting Process

Assuming a stronger constraint on the limiter, it is possible to show [1] that there hold a discrete version of the analytical properties (4.a-b). We substitute the constraint (COND.1) by the condition

$$0 \le \sigma_{ij}(\mathbf{c}) \begin{cases} \le 1 & \text{if } \nu_{ij} > 0, \\ < \infty & \text{otherwise.} \end{cases} \qquad \text{(COND.2)}$$

Proposition 1
Under the assumptions

- (COND.2) *on* $\sigma_{ij}(\mathbf{c})$*;*
- $c_i(0) \ge 0$ *for any* i*;*
- $\mathrm{g}(t, \mathbf{x}_{ij}) \ge 0$*;*

the FV approximate solution satisfies

$$0 \le c_i(t) \le M(t), \qquad \mathsf{T}_i \in \mathcal{T}_h.$$

Let us introduce a mesh-dependent norm and a scalar product respectively defined by

- $\|\mathbf{c}\|_{p,h} = \left(\sum\limits_{i=1}^{n_t} |\mathsf{T}_i| \, |c_i|^p \right)^{\frac{1}{p}}$,

- $\langle \gamma \mathbf{c}, v_n \rangle_h = \sum_{\mathsf{e}_{ij} \in \mathcal{E}'_h} \gamma \mathbf{c}|_{ij} \, \nu_{ij},$

where $\gamma \mathbf{c}|_{ij}$ is the *discrete trace*

$$\gamma \mathbf{c}|_{ij} = \begin{cases} \mathrm{g}(\cdot, \mathbf{x}_{ij}) & \text{if } \mathsf{e}_{ij} \subset \Gamma^-, \\ \mathcal{R}_i(\cdot, \mathbf{x}_{ij}; \mathbf{c}) & \text{otherwise.} \end{cases}$$

Proposition 2
Under the same assumptions of proposition 1 there holds that

$$\frac{d}{dt} \|\mathbf{c}\|_{1,h} + \langle \gamma \mathbf{c}, v_n \rangle_h = 0.$$

Let us introduce the following semi-norms and discrete scalar product

- $|\mathbf{c}|^2_{2,h,U} = \sum_{\mathsf{e}_{ij} \in \mathcal{E}^I_h} |\nu_{ij}| \, (c_i - c_j)^2$
- $|\mathbf{c}|^2_{2,h,R} = \sum_{\mathsf{e}_{ij} \in \mathcal{E}^I_h} 2(\nu^+_{ij} \sigma_{ij}(\mathbf{c}) + \nu^+_{ji} \sigma_{ji}(\mathbf{c}))(c_i - c_j)^2$
- $\langle \pi \mathbf{c}, v_n \rangle_h = \sum_{\mathsf{e}_{ij} \in \mathcal{E}'_h} \pi \mathbf{c}|_{ij} \, \nu_{ij}$

where $\pi \mathbf{c}|_{ij}$ is the *discrete border projection*

$$\pi \mathbf{c}|_{ij} = c_i$$

Proposition 3
Under the same assumptions of proposition 1, there holds that

$$\frac{d}{dt} \|\mathbf{c}\|^2_{2,h} + \left\langle (\gamma \mathbf{c})^2, v_n \right\rangle_h + A(\mathbf{c}) - B(\mathbf{c}) = 0$$

where

$$A(\mathbf{c}) = |\mathbf{c}|^2_{2,h,U} - |\mathbf{c}|^2_{2,h,R}, \qquad B(\mathbf{c}) = \left\langle ((\pi - \gamma)\mathbf{c})^2, v_n \right\rangle_h.$$

Let us notice that the term $+|\mathbf{c}|_{2,h,U}$ is related to the *upwind* numerical diffusion, while $-|\mathbf{c}|_{2,h,R}$ is a kind of numerical *anti*-diffusion related to the reconstruction. We can also remark that in general it is false that $A(\mathbf{c}) \geq 0$, i.e $|\mathbf{c}|^2_{2,h,U} \geq |\mathbf{c}|^2_{2,h,R}$, while there holds that the $B(\mathbf{c}) = \mathcal{O}\left(h^2\right)$.

Clearly, it is a highly desirable feature that $A(\mathbf{c})$ be non-negative and as small as possible, in order to reduce the difference $A(\mathbf{c}) - B(\mathbf{c})$, that is to control the numerical diffusion by the numerical anti-diffusion. Sufficient conditions can be formally stated that are capable of ensuring this feature as follows. Let us substitute the constraint (COND.2) by the stronger one:

$$0 \leq \sigma_{ij}(\mathbf{c}) \begin{cases} \leq 1/2 & \text{if } \nu_{ij} > 0, \\ < \infty & \text{otherwise.} \end{cases} \qquad \text{(COND.3)}$$

Then, the following proposition hold.

Proposition 4
Under the constraints

- (COND.3) *on* $\sigma_{ij}(\mathbf{c})$*;*
- $c_i(0) \geq 0$ *for any* i*;*
- $\mathrm{g}(t, \mathbf{x}_{ij}) \geq 0$*;*

the FV approximate solution satisfies

$$(i) \qquad |\mathbf{c}|^2_{2,h,a} \geq |\mathbf{c}|^2_{2,h,b}\,, \qquad \textit{that is} \qquad A(\mathbf{c}) \geq 0,$$

$$(ii) \qquad A = \mathcal{O}\left(h^2\right), \qquad \textit{on smooth solutions.}$$

7. Conclusions

The effects of both the first-order upwind dissipation and the polynomial reconstruction are expressed in terms of suitable mesh-dependent seminorms of the discrete solution field $\mathbf{c}$. In the case of a linear reconstruction algorithm the discrete version of (a), (b) and (c) does not hold; however, the simple limiting strategies (COND.1)–(COND.2) restore the discrete version of points (a), (b) and (c). The *anti-diffusive* effect of the recontruction process is outlined in a dimensionally independent way.

References

[1] E. Bertolazzi and G. Manzini. *Limiting strategies for polynomial reconsruction in Finite Volume Methods*. In preparation.

[2] P. G. Ciarlet. *The finite element method for elliptic problems* (North-Holland Publishing Company, Amsterdam, Holland, 1980).

[3] R. Eymard, T. Gallouët and R. Herbin. *Finite Volume methods*. In *Handbook of numerical analysis, Vol. VII* (North-Holland, Amsterdam, 2000), pages 713–1020.

[4] C. Hirsch. *Numerical Computation of Internal and External Flows* (J. Wiley & Sons Ltd., Baffins Lane, Chichester, West Sussex PO19 1UD, England, 1990).

[5] M. E. Hubbard. *Multidimensional slope limiters for MUSCL-type finite volume schemes on unstructured grids*. J. Comput. Phys., **155** (1999) 54–74.