SECOND-ORDER SLOPE LIMITERS FOR THE SIMULTANEOUS LINEAR ADVECTION OF (NOT SO) INDEPENDENT VARIABLES*

QUANG HUY TRAN†

Abstract. We propose a strategy to perform second-order enhancement using slope-limiters for the simultaneous linear advection of several scalar variables. Our strategy ensures a discrete min-max principle not only for each variable but also for any number of non-trivial combinations of them, which represent *control variables*. This problem arises in fluid mechanics codes using the Arbitrary Lagrange-Euler formalism, where the additional monotonicity property on control variables is required by physical considerations within the *remap* step.

Key words. Linear advection, min-max principle, slope reconstruction, Arbitrary Lagrange-Euler formalism

Subject classifications. 76M12, 65M06, 35L65

1. Introduction We are concerned with the design of a second-order scheme for the numerical solution of the system of linear advections

$$\partial_t \mathbf{\Psi} + u\partial_x \mathbf{\Psi} = \mathbf{0}, \tag{1.1}$$

where $u \in \mathbb{R}$ is a given velocity field and $\mathbf{\Psi} = (\psi^1, \ldots, \psi^P) \in \mathbb{R}_+^P$, with $P \in \mathbb{N}^*$, is the vector of positive unknowns. Although the components of $\mathbf{\Psi}$, called *main variables*, are independent from each other at the continuous level, our objective is to ensure a min-max principle at the discrete level not only for the main variables but also for a set of physically meaningful *control variables*

$$\mathbf{G}(\mathbf{\Psi}) = (G^1(\mathbf{\Psi}), \ldots, G^Q(\mathbf{\Psi})) \in \mathbb{R}^Q, \qquad Q \in \mathbb{N}^*. \tag{1.2}$$

The motivation for such a requirement usually comes from the context of the industrial application at hand and from the observation that, as a consequence of (1.1), the control quantities are also transported by the advection system

$$\partial_t \mathbf{G}(\mathbf{\Psi}) + u\partial_x \mathbf{G}(\mathbf{\Psi}) = \mathbf{0}. \tag{1.3}$$

In order to state the problem more accurately, let us recall some background material for the scalar linear advection

$$\partial_t \psi + u\partial_x \psi = 0, \tag{1.4}$$

in which the velocity u is assumed to be uniform and positive. Over a uniform grid with mesh-size Δx, the cells of which are indexed by i, and for a time-step Δt, we consider the explicit second-order update formula [9]

$$\widehat{\psi}_i = \psi_i - \lambda \left\{ \left[\psi_i + \tfrac{1-\lambda}{2} D_i^\psi \right] - \left[\psi_{i-1} + \tfrac{1-\lambda}{2} D_{i-1}^\psi \right] \right\}, \tag{1.5}$$

where D_i^ψ is the slope (multiplied by Δx) of ψ in cell i, and

$$\lambda = \frac{u\Delta t}{\Delta x} \tag{1.6}$$

*

†Département Mathématiques Appliquées, Institut Français du Pétrole, 1 et 4 avenue de Bois-Préau, 92852 Rueil-Malmaison Cedex, France, (Q-Huy.Tran@ifp.fr).

is the CFL ratio. It is well-known [4, 7] that if the slopes are suitably chosen via an appropriate limiter function

$$D_i^{\psi} = \widetilde{D}_i^{\psi} \equiv \Lambda(\psi_i - \psi_{i-1}, \psi_{i+1} - \psi_i), \tag{1.7}$$

such as minmod, van Leer, superbee... then there holds the discrete min-max principle

$$\widehat{\psi}_i \in \lfloor \psi_{i-1}, \psi_i \rceil \tag{1.8}$$

under the CFL condition $\lambda < 1$. We systematically use the notation $\lfloor \mathfrak{a}, \mathfrak{b} \rceil$ for the convex interval spanned by the real numbers $\mathfrak{a}$ and $\mathfrak{b}$.

Of course, we are going to apply scheme (1.5) to numerically solve system (1.1) component-wise. If the slopes are limited as in (1.7), i.e.,

$$D_i^{\psi^p} = \widetilde{D}_i^{\psi^p} \equiv \Lambda(\psi_i^p - \psi_{i-1}^p, \psi_{i+1}^p - \psi_i^p) \tag{1.9}$$

for $1 \leq p \leq P$, then there holds discrete min-max principle component-wise

$$\widehat{\psi}_i^p \in \lfloor \psi_{i-1}^p, \psi_i^p \rceil. \tag{1.10}$$

As for the control variables, which are necessarily computed as

$$G_i^q = G^q(\mathbf{\Psi}_i), \qquad \widehat{G}_i^q = G^q(\widehat{\mathbf{\Psi}}_i) \tag{1.11}$$

for $1 \leq q \leq Q$, there is no reason that we should have the desired min-max principle

$$\widehat{G}_i^q \in \lfloor G_{i-1}^q, G_i^q \rceil, \tag{1.12}$$

insofar as the components of $\mathbf{\Psi}$ do not "see each other."

The min-max principle on control variables is a major challenge in many industrial fluid mechanics codes using an ALE (Arbitrary Lagrange-Euler) method [2, 3], the remap phase of which consists in simultaneously advecting several supposedly independent variables. Such a requirement is essential for robustness. However, except for a partially successful attempt by VanderHeyden and Kashiwa [10] for a restricted setting of the fraction problem, we do not have knowledge of any thoroughly satisfactory solution. The present contribution demonstrates that the component-wise limitation (1.9) can be actually replaced by a more general framework

$$D_i^{\psi^p} = \Lambda^p(\mathbf{\Psi}_{i-1}, \mathbf{\Psi}_i, \mathbf{\Psi}_{i+1}) \tag{1.13}$$

which does guarantee (1.10) and (1.12) under the same CFL condition $\lambda < 1$. This novel procedure can be extended to the case of a space-dependent velocity field $u = u(x)$, the sign of which is not necessarily constant. From a practical point of view, the new slopes will be obtained from the old ones, computed by (1.9), through a projection mechanism which creates the opportunity for the various (main and control) variables to see each other. This projection mechanism is optimally designed in order for the new slopes to be as "close" as possible to the old slopes in some sense, so that sharp profiles can still be captured.

In order to convey the geometric insights that are at the root of the seemingly complex algebraic formalism of this work, we focus on two simplest but most important examples encountered in the context of Euler-like fluid models: the sum problem §2 for a flame model [1] and the fraction problem §3 for a two-phase flow model [2]. Section 4 is devoted to the general problem, along with some examples selected from real-life applications.

2. The sum problem We are interested in the densities of two species, say, CH_4 and CO_2, as well as in their sum which represent the carbon tracer. Let us put

$$\mathbf{\Psi} = (\alpha, \beta) \in \mathbb{R}_+^2, \qquad G(\mathbf{\Psi}) = \alpha + \beta \in \mathbb{R}_+. \tag{2.1}$$

2.1. Uniform velocity Assume $u(x) = u > 0$. From a time level to the next one, the update formulae for (α, β) are

$$\widehat{\alpha}_i = \alpha_i - \lambda\{[\alpha_i + \tfrac{1-\lambda}{2} D_i^\alpha] - [\alpha_{i-1} + \tfrac{1-\lambda}{2} D_{i-1}^\alpha]\} \tag{2.2a}$$

$$\widehat{\beta}_i = \beta_i - \lambda\{[\beta_i + \tfrac{1-\lambda}{2} D_i^\beta] - [\beta_{i-1} + \tfrac{1-\lambda}{2} D_{i-1}^\beta]\}, \tag{2.2b}$$

where λ is defined in (1.6). Consider the *initial slopes*

$$\widetilde{D}_i^\alpha = \Lambda(\alpha_i - \alpha_{i-1}, \alpha_{i+1} - \alpha_i), \qquad \widetilde{D}_i^\beta = \Lambda(\beta_i - \beta_{i-1}, \beta_{i+1} - \beta_i), \tag{2.3}$$

inspired from the scalar case (1.7) and computed component-wise via a standard limiter function Λ, such as minmod, van Leer, superbee or ultrabee [4, 7]. The following Lemma recalls a useful property. We use the notations $r^- = \min(r, 0)$ and $r^+ = \max(r, 0)$ for the negative and positive part of any real number r.

LEMMA 2.1. *If the slopes (D_j^α, D_j^β) in (2.2) satisfy*

$$[\widetilde{D}_j^\alpha]^- \le D_j^\alpha \le [\widetilde{D}_j^\alpha]^+, \qquad [\widetilde{D}_j^\beta]^- \le D_j^\beta \le [\widetilde{D}_j^\beta]^+ \tag{2.4}$$

for $j = i-1$ and $j = i$, then $\widehat{\alpha}_i \in \lfloor \alpha_{i-1}, \alpha_i \rceil$ and $\widehat{\beta}_i \in \lfloor \beta_{i-1}, \beta_i \rceil$.

Proof. This is a consequence of Sweby's analysis [7], by which an appropriate choice of the limiter function Λ allows one to express $\widehat{\alpha}_i$ as a convex combination of α_{i-1} and α_i (likewise for $\widehat{\beta}_i$). Conditions (2.4) say that the new slopes must be of the same sign as the old ones, while having smaller absolute values. □

The check-sum variable G is computed by $G_i = \alpha_i + \beta_i$ and $\widehat{G}_i = \widehat{\alpha}_i + \widehat{\beta}_i$. It will be a mistake to take it for granted that the min-max principles on α and β always imply that on G: this is true only when the min (resp. max) of the sum is equal to the sum of the min values (resp. max values), which means that α and β both increase or both decrease from $i-1$ to i, as highlighted by the following Proposition.

PROPOSITION 2.2. *If $(\alpha_i - \alpha_{i-1})(\beta_i - \beta_{i-1}) \ge 0$ and if the slopes (D_j^α, D_j^β) satisfy (2.4) for $j = i-1$ and $j = i$, then $\widehat{G}_i \in \lfloor G_{i-1}, G_i \rceil$.*

Proof. In the quarter-plane $(\alpha, \beta) \in \mathbb{R}_+ \times \mathbb{R}_+$, let us depict the points

$$M_{i-1} = (\alpha_{i-1}, \beta_{i-1}), \quad M_i = (\alpha_i, \beta_i), \quad \widehat{M}_i = (\widehat{\alpha}_i, \widehat{\beta}_i). \tag{2.5}$$

as in Figure 2.1. The min-max principles $\widehat{\alpha}_i \in \lfloor \alpha_{i-1}, \alpha_i \rceil$ and $\widehat{\beta}_i \in \lfloor \beta_{i-1}, \beta_i \rceil$, which follow from Lemma 2.1, amount to saying that $\widehat{M}_i$ belongs to the rectangle $\mathscr{R}_i$ whose opposite vertices are M_{i-1} and M_i and whose sides are parallel to the horizontal and vertical axes. Draw the lines $\mathfrak{G}_{i-1}$ and $\mathfrak{G}_i$ defined by $\alpha + \beta = G_{i-1}$ and $\alpha + \beta = G_i$.

If $(\alpha_i - \alpha_{i-1})(\beta_i - \beta_{i-1}) \ge 0$, then the rectangle $\mathscr{R}_i$ is entirely included in the strip defined by the parallel lines $\mathfrak{G}_{i-1}$ and $\mathfrak{G}_i$. Therefore, the isoline of $\alpha + \beta$ passing through $\widehat{M}_i$ lies between $\mathfrak{G}_{i-1}$ and $\mathfrak{G}_i$, which is algebraically equivalent to $\widehat{G}_i \in \lfloor G_{i-1}, G_i \rceil$.

If $(\alpha_i - \alpha_{i-1})(\beta_i - \beta_{i-1}) < 0$, the lines $\mathfrak{G}_{i-1}$ and $\mathfrak{G}_i$ cut the rectangle $\mathscr{R}_i$ into three pieces, and it may happen that $\widehat{M}_i$ lies outside the strip, which violates the desired min-max principle. □

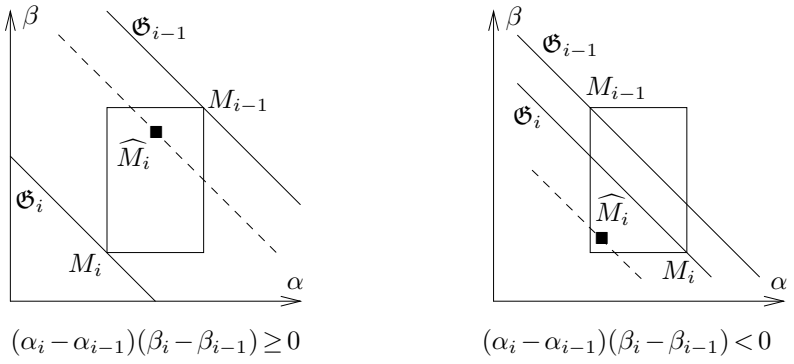

FIG. 2.1. *Geometric analysis of the min-max principle for the sum problem.*

To know what should be done for the case $(\alpha_i-\alpha_{i-1})(\beta_i-\beta_{i-1})<0$, we introduce

$$G_i^m=\min\{G_{i-1},G_i\},\qquad G_i^M=\max\{G_{i-1},G_i\},\tag{2.6}$$

and seek sufficient conditions in terms of (D^α,D^β) so that $G_i^m\leq\widehat{G}_i\leq G_i^M$ at a given cell i. Unless otherwise indicated, it is assumed that $\lambda<1$.

LEMMA 2.3. *For a given cell i, if*

$$\tfrac{2}{\lambda}(G_i\quad-G_i^M)\leq\quad D_i^\alpha+D_i^\beta\quad\leq\qquad\tfrac{2}{\lambda}(G_i\quad-G_i^m),\tag{2.7a}$$

$$-\tfrac{2}{1-\lambda}(G_{i-1}-G_i^m)\leq D_{i-1}^\alpha+D_{i-1}^\beta\leq-\tfrac{2}{1-\lambda}(G_{i-1}-G_i^M),\tag{2.7b}$$

then $G_i^m\leq\widehat{G}_i\leq G_i^M$.

Proof. Subtracting the convex decomposition $G_i^m=(1-\lambda)G_i^m+\lambda G_i^m$ to the sum of (2.2a) and (2.2b), we obtain

$$\widehat{G}_i-G_i^m=(1-\lambda)[G_i-G_i^m-\tfrac{\lambda}{2}(D_i^\alpha+D_i^\beta)]+\lambda[G_{i-1}-G_i^m+\tfrac{1-\lambda}{2}(D_{i-1}^\alpha+D_{i-1}^\beta)].\tag{2.8}$$

To ensure $\widehat{G}_i-G_i^m\geq 0$, we split the right-hand side into two parts and impose positivity to each summand. This leads to the right part of (2.7a) and the left part of (2.7b). We proceed similarly to impose negativity to $\widehat{G}_i-G_i^M$. ∎

The benefit of this splitting approach lies in the fact that the resulting conditions (2.7) are local: they do not couple the slopes at cell i with those at cell $i-1$, thus giving rise to a tractable procedure. It could be legitimately feared that imposing positivity separately in (2.8) yields too strong conditions which might deteriorate accuracy. The miracle is yet that accuracy is preserved at a very good level, as shown by numerical results.

We are now in a position to formulate the new limitation procedure, which ensures the min-max principle for all cells.

THEOREM 2.4. *Given an initial choice $(\widetilde{D}_i^\alpha,\widetilde{D}_i^\beta)$ in accordance with (2.3), let $\mathscr{G}_i\subset\mathbb{R}^2$ be the set of all pairs (D_i^α,D_i^β) subject to the 6 linear inequality constraints*

$$[\widetilde{D}_i^\alpha]^-\leq D_i^\alpha\leq[\widetilde{D}_i^\alpha]^+,\quad[\widetilde{D}_i^\beta]^-\leq D_i^\beta\leq[\widetilde{D}_i^\beta]^+,\quad\mathfrak{m}_i\leq D_i^\alpha+D_i^\beta\leq\mathfrak{M}_i,\tag{2.9}$$

with

$$\mathfrak{m}_i = \max\{\tfrac{2}{\lambda}[G_i - G_{i-1}]^-\,;\, \tfrac{2}{1-\lambda}[G_{i+1} - G_i]^-\} \tag{2.10a}$$

$$\mathfrak{M}_i = \min\{\tfrac{2}{\lambda}[G_i - G_{i-1}]^+\,;\, \tfrac{2}{1-\lambda}[G_{i+1} - G_i]^+\}. \tag{2.10b}$$

For all $i \in \mathbb{Z}$, define

$$(D_i^\alpha, D_i^\beta) = \begin{cases} (\widetilde{D}_i^\alpha, \widetilde{D}_i^\beta) & \text{if } \widetilde{D}_i^\alpha \widetilde{D}_i^\beta > 0 \\ \Pi_{\mathscr{G}_i}(\widetilde{D}_i^\alpha, \widetilde{D}_i^\beta) & \text{otherwise} \end{cases} \tag{2.11}$$

where $\Pi_{\mathscr{G}_i}(.)$ denotes the projection onto the convex set $\mathscr{G}_i \subset \mathbb{R}^2$. Then, we have the min-max principles

$$\widehat{\alpha}_i \in \lfloor \alpha_{i-1}, \alpha_i \rceil, \qquad \widehat{\beta}_i \in \lfloor \beta_{i-1}, \beta_i \rceil, \qquad \widehat{G}_i \in \lfloor G_{i-1}, G_i \rceil, \tag{2.12}$$

at every cell i when updating (α, β) with scheme (2.2).

Proof. The set $\mathscr{G}_i$ is obviously convex not empty because it contains $(0,0)$. Therefore, definition (2.11) makes sense. Note that its shape depends on $(\widetilde{D}_i^\alpha, \widetilde{D}_i^\beta)$.

At a fixed cell i, if $\widetilde{D}_i^\alpha \widetilde{D}_i^\beta > 0$, we necessarily have $(\alpha_i - \alpha_{i-1})(\beta_i - \beta_{i-1}) > 0$ on the grounds of the properties of standard limiter functions. According to Proposition 2.2, the default values $(D_i^\alpha, D_i^\beta) = (\widetilde{D}_i^\alpha, \widetilde{D}_i^\beta)$ are suitable. If $\widetilde{D}_i^\alpha \widetilde{D}_i^\beta \leq 0$, we are going to check conditions (2.7). From (2.9) and (2.10), we infer that

$$\tfrac{2}{\lambda}[G_i - G_{i-1}]^- \leq \mathfrak{m}_i \leq D_i^\alpha + D_i^\beta \leq \mathfrak{M}_i \leq \tfrac{2}{\lambda}[G_i - G_{i-1}]^+. \tag{2.13}$$

We claim that $[G_i - G_{i-1}]^- = G_i - G_i^M$ and $[G_i - G_{i-1}]^+ = G_i - G_i^m$. To see this, we simply have to distinguish two cases $G_{i-1} \leq G_i$ and $G_{i-1} > G_i$. This establishes (2.7a).

If $\widetilde{D}_{i-1}^\alpha \widetilde{D}_{i-1}^\beta > 0$, we also necessarily have $(\alpha_i - \alpha_{i-1})(\beta_i - \beta_{i-1}) > 0$ thanks to the properties of standard limiter functions. By virtue of Proposition 2.2, we conclude that there is no need to check (2.7b). If $\widetilde{D}_{i-1}^\alpha \widetilde{D}_{i-1}^\beta \leq 0$, we write (2.9) and (2.10) for cell $i-1$ and observe that

$$\tfrac{2}{1-\lambda}[G_i - G_{i-1}]^- \leq \mathfrak{m}_{i-1} \leq D_{i-1}^\alpha + D_{i-1}^\beta \leq \mathfrak{M}_{i-1} \leq \tfrac{2}{1-\lambda}[G_i - G_{i-1}]^+, \tag{2.14}$$

and once again argue that $[G_i - G_{i-1}]^- = G_i - G_i^M$ and $[G_i - G_{i-1}]^+ = G_i - G_i^m$ to derive (2.7b). □

The coding of the projection operator $\Pi_{\mathscr{G}_i}$ in this problem can be made efficient through explicit formulae. Figure 2.2 illustrates a few situations for a locally increasing or decreasing behavior of G. Note that if a local extremum occurs, i.e., $(G_{i-1} - G_i)(G_{i+1} - G_i) > 0$, by (2.10) we have $\mathfrak{m}_i = \mathfrak{M}_i = 0$, hence $D_i^G = D_i^\alpha + D_i^\beta = 0$. This testifies to a clipping mechanism on G in the proposed procedure.

2.2. Variable velocity field The velocities $u_{i+1/2} = u(x_{i+1/2})$ are given at the edges. The advection equation $\partial_t \psi + u \partial_x \psi = 0$ is discretized by the explicit scheme

$$\begin{aligned} \widehat{\psi}_i = \psi_i &- \lambda_{i-1/2}^- \tfrac{1-|\lambda|_i}{2} D_i^\psi - \lambda_{i-1/2}^+ \left\{ \psi_i - \left[\psi_{i-1} + \tfrac{1-|\lambda|_{i-1}}{2} D_{i-1}^\psi \right] \right\} \\ &- \lambda_{i+1/2}^+ \tfrac{1-|\lambda|_i}{2} D_i^\psi + \lambda_{i+1/2}^- \left\{ \psi_i - \left[\psi_{i+1} - \tfrac{1-|\lambda|_{i+1}}{2} D_{i+1}^\psi \right] \right\}, \end{aligned} \tag{2.15}$$

where

$$\lambda_{i\pm 1/2} = \frac{u_{i\pm 1/2} \Delta t}{\Delta x}, \qquad |\lambda|_i = \lambda_{i-1/2}^+ - \lambda_{i+1/2}^-. \tag{2.16}$$

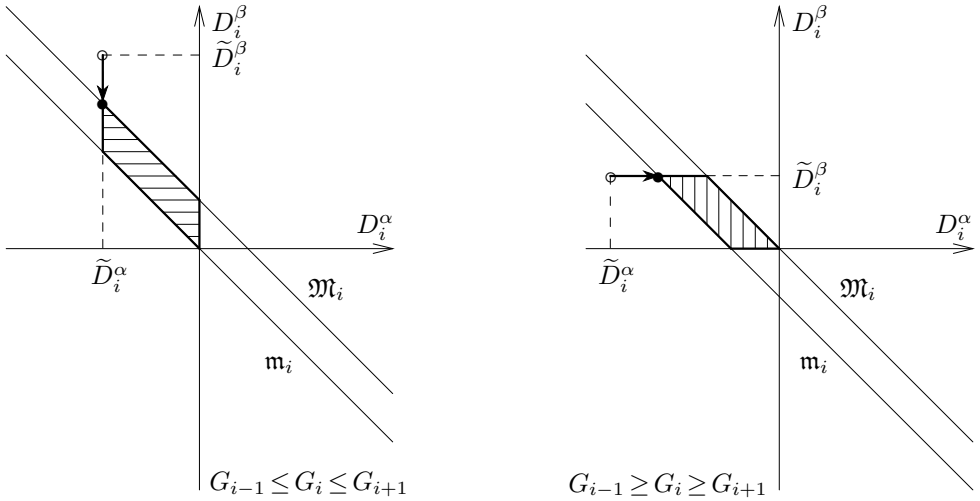

FIG. 2.2. *Projection of* $(\widetilde{D}_i^\alpha, \widetilde{D}_i^\beta)$ *onto the convex set* $\mathscr{G}_i$ *for the sum problem.*

Such a scheme stems from considerations about conservativity in the remap phase of the ALE method. We recall [5] that it can only be quasi second-order for a variable velocity field.

For clarity of language, let us give names to the various situations that may occur depending on the sign configuration $S(i)$ in the neighborhood of cell i.

I.	$u_{i-1/2} > 0$ and $u_{i+1/2} \ge 0$	(left to right propagation);
II.	$u_{i-1/2} \le 0$ and $u_{i+1/2} < 0$	(right to left propagation);
III.	$u_{i-1/2} \le 0$ and $u_{i+1/2} \ge 0$	(source);
IV.	$u_{i-1/2} > 0$ and $u_{i+1/2} < 0$	(sink).

The min-max principle on ψ reads

$$\widehat{\psi}_i \in \begin{cases} \lfloor \psi_{i-1}, \psi_i \rceil & \text{if } S(i) = \text{I}, \\ \lfloor \psi_{i+1}, \psi_i \rceil & \text{if } S(i) = \text{II}, \\ \lfloor \psi_{i-1}, \psi_i, \psi_{i+1} \rceil & \text{if } S(i) = \text{III} \cup \text{IV}. \end{cases} \tag{2.17}$$

The aim of the game is the same as before: we transport $\psi = \alpha$ and $\psi = \beta$ by (2.15) but require the min-max principle (2.17) on $\psi = \alpha, \beta$ and G. Of course, we rely on the kind on analysis as in the uniform velocity case, although the discussion becomes much more involved.

For simplicity, we make additional assumptions in order to have statements similar to the uniform case.

1. The CFL condition is about half the previous one. For all cell i, we have

$$|\lambda_{i-1/2}| + |\lambda_{i+1/2}| < 1. \tag{2.18}$$

2. The standard limiter function Λ used to compute the initial slopes (2.3) at cell i is of *strength* lesser than $2/(1-|\lambda|_i)$, i.e.,

$$|\Lambda(d_{i-1/2}, d_{i+1/2})| \le \frac{2}{1-|\lambda|_i} \min\{|d_{i-1/2}|, |d_{i+1/2}|\}. \tag{2.19}$$

This rules out ultrabee, but authorizes minmod, van Leer, superbee and even hyperbee based on $|\lambda|_i$.

3. There is no sequence of *source-sink* (C–D) or *sink-source* (D–C) configuration over two consecutive cells. Put another way,

$$\nexists i \in \mathbb{Z} \mid \lambda_{i-1/2}\lambda_{i+1/2} < 0 \quad \text{and} \quad \lambda_{i+1/2}\lambda_{i+3/2} < 0. \tag{2.20}$$

Such a "saw-tooth" sequence can be avoided by refining the mesh sufficiently, provided that the velocity field $u(x)$ depends continuously on x.

In preparation for Theorem 2.5, we set

$$\Phi_i = \min\{\tfrac{2}{|\lambda|_i}, \tfrac{2}{1-|\lambda|_i}\} \tag{2.21}$$

and introduce the local bounds

$$\begin{aligned} G_i^m &= \mathbb{1}_{\{S(i)=\mathrm{I}\}} \ \min\{G_{i-1}, G_i\} && \text{(2.22a)}\\ &\quad + \mathbb{1}_{\{S(i)=\mathrm{II}\}} \min\{G_{i+1}, G_i\} + \mathbb{1}_{\{S(i)=\mathrm{III}\cup\mathrm{IV}\}} \min\{G_{i-1}, G_i, G_{i+1}\} \\ G_i^M &= \mathbb{1}_{\{S(i)=\mathrm{I}\}} \ \max\{G_{i-1}, G_i\} && \text{(2.22b)}\\ &\quad + \mathbb{1}_{\{S(i)=\mathrm{II}\}} \max\{G_{i+1}, G_i\} + \mathbb{1}_{\{S(i)=\mathrm{III}\cup\mathrm{IV}\}} \max\{G_{i-1}, G_i, G_{i+1}\}, \end{aligned}$$

where $\mathbb{1}_{\{.\}}$ is the characteristic function. Once all the G^m and G^M have been computed over the domain, we consider the set $\mathscr{G}_i$ of all pairs (D_i^α, D_i^β) that satisfy:

- For case I (left-to-right propagation)

$$[\widetilde{D}_i^\alpha]^- \le D_i^\alpha \le [\widetilde{D}_i^\alpha]^+, \quad [\widetilde{D}_i^\beta]^- \le D_i^\beta \le [\widetilde{D}_i^\beta]^+, \quad \mathfrak{m}_i \le D_i^\alpha + D_i^\beta \le \mathfrak{M}_i, \tag{2.23}$$

with

$$\begin{aligned} \mathfrak{m}_i &= \Phi_i \max\{G_i^m - G_i \ \ ; G_i \ - G_{i+1}^M\} && \text{(2.24a)}\\ \mathfrak{M}_i &= \Phi_i \min\{G_i \ - G_{i+1}^m ; G_i^M - G_i\}. && \text{(2.24b)} \end{aligned}$$

- For case II (right-to-left propagation)

$$[\widetilde{D}_i^\alpha]^- \le D_i^\alpha \le [\widetilde{D}_i^\alpha]^+, \quad [\widetilde{D}_i^\beta]^- \le D_i^\beta \le [\widetilde{D}_i^\beta]^+, \quad \mathfrak{m}_i \le D_i^\alpha + D_i^\beta \le \mathfrak{M}_i, \tag{2.25}$$

with

$$\begin{aligned} \mathfrak{m}_i &= \Phi_i \max\{G_i^m - G_i \ \ ; G_i \ - G_{i-1}^M\} && \text{(2.26a)}\\ \mathfrak{M}_i &= \Phi_i \min\{G_i \ - G_{i-1}^m ; G_i^M - G_i\}. && \text{(2.26b)} \end{aligned}$$

- For case III (source)

$$[\widetilde{D}_i^\alpha]^- \le D_i^\alpha \le [\widetilde{D}_i^\alpha]^+, \quad [\widetilde{D}_i^\beta]^- \le D_i^\beta \le [\widetilde{D}_i^\beta]^+, \quad \mathfrak{m}_i \le D_i^\alpha + D_i^\beta \le \mathfrak{M}_i, \tag{2.27}$$

with

$$\begin{aligned} \mathfrak{m}_i &= \Phi_i \max\{G_i - G_{i-1}^M ; G_i - G_i^M ; G_i^m - G_i \ ; G_{i+1}^m - G_i\} && \text{(2.28a)}\\ \mathfrak{M}_i &= \Phi_i \min\{G_i - G_{i-1}^m ; G_i - G_i^m \ ; G_i^M - G_i ; G_{i+1}^M - G_i\}. && \text{(2.28b)} \end{aligned}$$

- For case IV (sink), $\mathscr{G}_i = \mathbb{R}^2$.

THEOREM 2.5. *Given an initial choice $(\widetilde{D}_i^\alpha,\widetilde{D}_i^\beta)$ in accordance with (2.3), (2.19), let $\mathscr{G}_i\subset\mathbb{R}^2$ be the convex set introduced above. For all $i\in\mathbb{Z}$, define*

$$(D_i^\alpha,D_i^\beta)=\begin{cases}(\widetilde{D}_i^\alpha,\widetilde{D}_i^\beta) & \text{if } \widetilde{D}_i^\alpha\widetilde{D}_i^\beta>0\\ \Pi_{\mathscr{G}_i}(\widetilde{D}_i^\alpha,\widetilde{D}_i^\beta) & \text{otherwise}\end{cases} \tag{2.29}$$

where $\Pi_{\mathscr{G}_i}(.)$ denotes the projection onto the convex set $\mathscr{G}_i\subset\mathbb{R}^2$. Then, under assumptions (2.18) and (2.20), we have the min-max principle (2.17) for $\psi=\alpha,\beta$ and G at every cell i when updating (α,β) with scheme (2.15).

Proof. Since the proof is lengthy and relatively tedious, we are going to sketch out its beginning in order for the readers to grasp the key ideas. Applying the update formula (2.15) to $\psi=\alpha$ and β, then adding the equations together and substracting by G_i^m and G_i^M yields

$$\widehat{G}_i-G_i^m=A_i^m+B_{i-1}^m+C_{i+1}^m,\qquad \widehat{G}_i-G_i^M=A_i^M+B_{i-1}^M+C_{i+1}^M, \tag{2.30}$$

with

$$A_i^m=(1-\lambda_{i-1/2}^+ +\lambda_{i+1/2}^-)(G_i-G_i^m)-(\lambda_{i+1/2}^+ +\lambda_{i-1/2}^-)\tfrac{1-|\lambda|_i}{2}D_i^G \tag{2.31a}$$

$$B_{i-1}^m=\lambda_{i-1/2}^+(G_{i-1}-G_i^m)+\lambda_{i-1/2}^+\tfrac{1-|\lambda|_{i-1}}{2}D_{i-1}^G \tag{2.31b}$$

$$C_{i+1}^m=\lambda_{i+1/2}^-(G_{i+1}-G_i^m)+\lambda_{i+1/2}^-\tfrac{1-|\lambda|_{i+1}}{2}D_{i+1}^G, \tag{2.31c}$$

and

$$A_i^M=(1-\lambda_{i-1/2}^+ +\lambda_{i+1/2}^-)(G_i-G_i^M)-(\lambda_{i+1/2}^+ +\lambda_{i-1/2}^-)\tfrac{1-|\lambda|_i}{2}D_i^G \tag{2.32a}$$

$$B_{i-1}^M=\lambda_{i-1/2}^+(G_{i-1}-G_i^M)+\lambda_{i-1/2}^+\tfrac{1-|\lambda|_{i-1}}{2}D_{i-1}^G \tag{2.32b}$$

$$C_{i+1}^M=\lambda_{i+1/2}^-(G_{i+1}-G_i^M)+\lambda_{i+1/2}^-\tfrac{1-|\lambda|_{i+1}}{2}D_{i+1}^G, \tag{2.32c}$$

using the shorthand notation $D^G=D^\alpha+D^\beta$. In conformity with the splitting philosophy already explained for the uniform velocity case, we separately impose

$$A_i^m\geq 0,\qquad B_{i-1}^m\geq 0,\qquad C_{i+1}^m\geq 0 \tag{2.33a}$$

$$A_i^M\leq 0,\qquad B_{i-1}^M\leq 0,\qquad C_{i+1}^M\leq 0. \tag{2.33b}$$

We then express (2.33) in terms of D^G according to the sign configuration. In case I (resp. II), we drop out the identically vanishing and useless inequalties on $C_{i+1}^{m,M}$ (resp. $B_{i-1}^{m,M}$) and we shift index for the inequalties on $B_{i-1}^{m,M}$ (resp. $C_{i+1}^{m,M}$) in order to ensure the min-max principle at the "receiving" neighbor $i+1$ (resp. $i-1$). In case III, we have to keep all the conditions and shift index for them, because a source does have an influence on two receiving neighbors. In case IV, there is no need to change D_i^G because a sink does not have any influence on its neighbors and the min-max principle at a sink is actually ensured by conditions imposed to the two neighbors. ∎

Despite its apparent complexity, this procedure lends itself very well to numerical implementation. Instead of finding the image of the projection by hands, we can resort to a subroutine for quadratic minimization under linear inequality constraints. This will be addressed in §4.

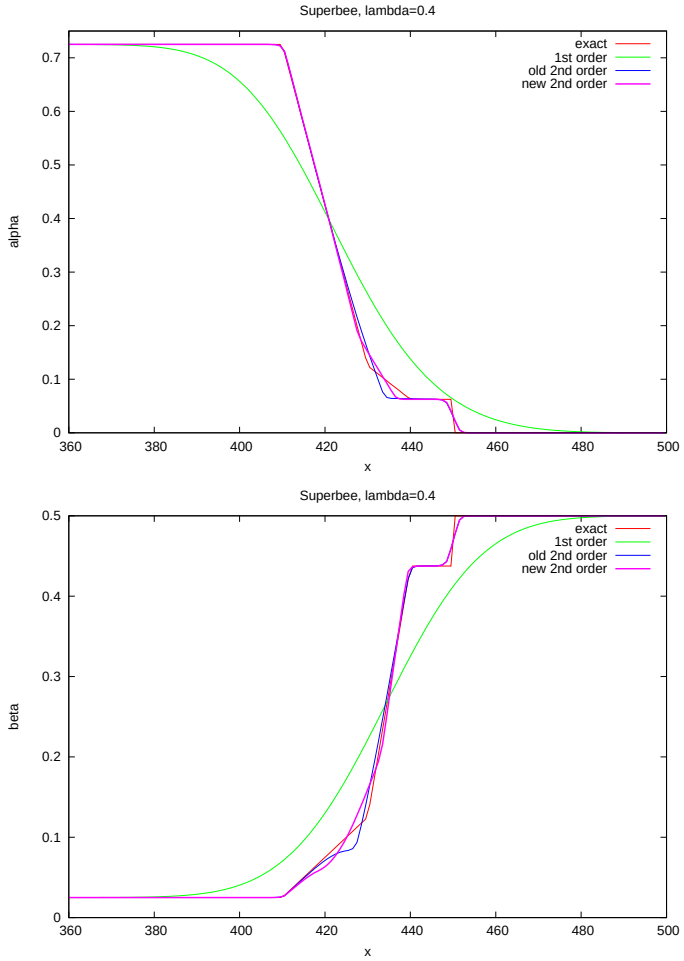

FIG. 2.3. *Main variables α (upper panel) and β (lower panel) for the sum problem.*

REMARK 2.6. *It can be shown that in the (D_i^α, D_i^β)-plane, the set of points defined by the inequalities $\mathfrak{m}_i \le D_i^\alpha + D_i^\beta \le \mathfrak{M}_i$ in cases* I, II *and* III *always contain the strip*

$$\max\{[G_i - G_{i-1}]^-, [G_{i+1} - G_i]^-\} \le \frac{D_i^\alpha + D_i^\beta}{\Phi_i} \le \min\{[G_i - G_{i-1}]^+, [G_{i+1} - G_i]^+\}.$$

Therefore, if we accept to project onto a smaller convex set, it is possible to find the new slopes by explicit formulae. The price to be paid for is a larger amount of dissipation.

2.3. Numerical results In Figures 2.3 and 2.4, we compare the results of 3 different schemes and the exact solution for an experiment over a positive velocity field. The initial data have been tailored so that α is decreasing and β is increasing, therefore we have $(\alpha_i - \alpha_{i-1})(\beta_i - \beta_{i-1}) \le 0$ for all cell i in the domain. The curves for the first-order scheme are very much smeared out. Those for the two second-order schemes are in very good agreement with the exact solution. What we mean by "old

second-order" is the scheme with the initial slopes, computed component-wise. Of course, the "new second-order" is endowed with our coupling device for the slopes.

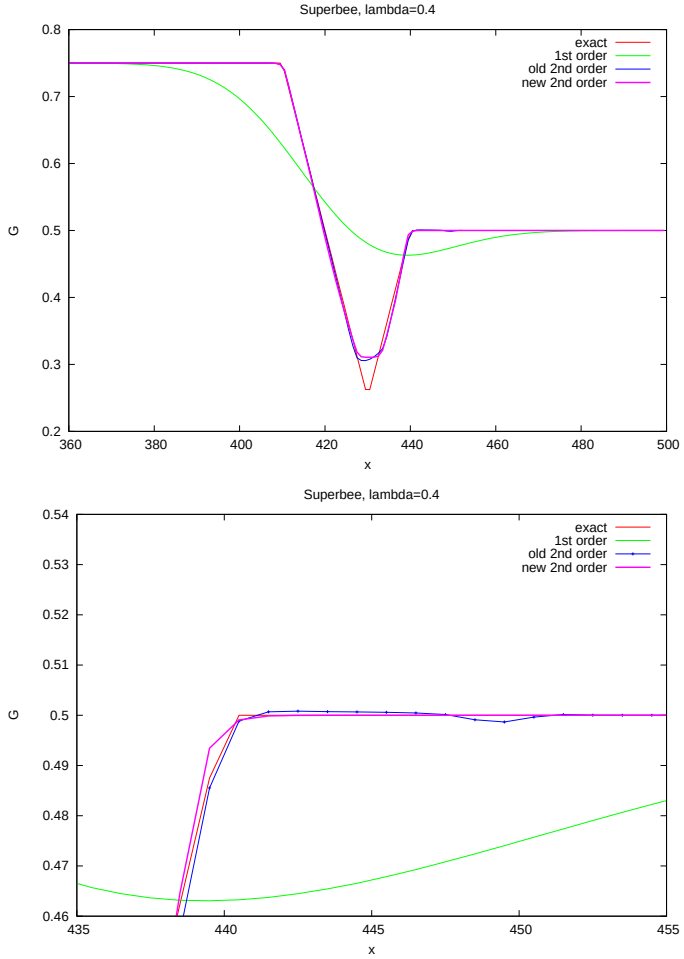

FIG. 2.4. *Control variable $G=\alpha+\beta$ for the sum problem.*

A close inspection of the curves reveals that in the vicinity of $x=445$, the old second-order scheme exhibits small spurious oscillations on the control variable G, as evidenced by the close-up in the lower panel of Figure 2.4. As far as the new second-order is concerned, there is no violation of the min-max principle.

3. The fraction problem We now turn to the transport of a total density and a partial density, the quotient of the latter by the former being a mass fraction. More specifically, we put

$$\mathbf{\Psi}=(\rho,\kappa)\in\mathbb{R}^2_+, \qquad Y(\mathbf{\Psi})=\frac{\kappa}{\rho}\in[0,1]. \tag{3.1}$$

3.1. Uniform velocity From a time level to the next one, the update formulae for (ρ,κ) are

$$\widehat{\rho}_i = \rho_i - \lambda\{[\rho_i + \tfrac{1-\lambda}{2}D_i^\rho] - [\rho_{i-1} + \tfrac{1-\lambda}{2}D_{i-1}^\rho]\} \tag{3.2a}$$

$$\widehat{\kappa}_i = \kappa_i - \lambda\{[\kappa_i + \tfrac{1-\lambda}{2}D_i^\kappa] - [\kappa_{i-1} + \tfrac{1-\lambda}{2}D_{i-1}^\kappa]\}, \tag{3.2b}$$

where λ is defined in (1.6). Consider the *initial slopes*

$$\widetilde{D}_i^\rho = \Lambda(\rho_i - \rho_{i-1}, \rho_{i+1} - \rho_i), \qquad \widetilde{D}_i^\kappa = \Lambda(\kappa_i - \kappa_{i-1}, \kappa_{i+1} - \kappa_i). \tag{3.3}$$

For the same reasons as in Lemma 2.1, we have the following result.

LEMMA 3.1. *If the slopes (D_j^ρ, D_j^κ) in (3.2) satisfy*

$$[\widetilde{D}_j^\rho]^- \le D_j^\rho \le [\widetilde{D}_j^\rho]^+, \qquad [\widetilde{D}_j^\kappa]^- \le D_j^\rho \le [\widetilde{D}_j^\kappa]^+ \tag{3.4}$$

for $j = i-1$ and $j = i$, then $\widehat{\rho}_i \in \lfloor \rho_{i-1}, \rho_i \rceil$ and $\widehat{\kappa}_i \in \lfloor \kappa_{i-1}, \kappa_i \rceil$.

The control fraction Y is computed by $Y_i = \kappa_i/\rho_i$ and $\widehat{Y}_i = \widehat{\kappa}_i/\widehat{\rho}_i$. Contrary to the intuitive feeling, it will be erroneous to think that carrying out the slope reconstruction on ρ and Y solves the problem. Indeed, the min (resp. max) of a product is not the product of the min values (resp. max values).

PROPOSITION 3.2. *If $(\rho_i - \rho_{i-1})(\kappa_i - \kappa_{i-1}) \le 0$ and if the slopes (D_j^ρ, D_j^κ) satisfy (3.4) for $j = i-1$ and $j = i$, then $\widehat{Y}_i \in \lfloor Y_{i-1}, Y_i \rceil$.*

Proof. In the quarter-plane $(\rho,\kappa) \in \mathbb{R}_+ \times \mathbb{R}_+$, let us depict the points

$$M_{i-1} = (\rho_{i-1}, \kappa_{i-1}), \quad M_i = (\rho_i, \kappa_i), \quad \widehat{M}_i = (\widehat{\rho}_i, \widehat{\kappa}_i). \tag{3.5}$$

as in Figure 3.1. The min-max principles $\widehat{\rho}_i \in \lfloor \rho_{i-1}, \rho_i \rceil$ and $\widehat{\kappa}_i \in \lfloor \kappa_{i-1}, \kappa_i \rceil$, which follow from Lemma 3.1, amount to saying that $\widehat{M}_i$ belongs to the rectangle $\mathscr{R}_i$ whose opposite vertices are M_{i-1} and M_i and whose sides are parallel to the horizontal and vertical axes. Draw the lines $\mathfrak{Y}_{i-1}$ and $\mathfrak{Y}_i$ defined by $\kappa/\rho = Y_{i-1}$ and $\kappa/\rho = Y_i$.

If $(\rho_i - \rho_{i-1})(\kappa_i - \kappa_{i-1}) \le 0$, then the rectangle $\mathscr{R}_i$ is entirely included in the cone of lines defined by the rays $\mathfrak{Y}_{i-1}$ and $\mathfrak{Y}_i$. Therefore, the isoline of κ/ρ passing through $\widehat{M}_i$ lies between $\mathfrak{Y}_{i-1}$ and $\mathfrak{Y}_i$, which is algebraically equivalent to $\widehat{Y}_i \in \lfloor Y_{i-1}, Y_i \rceil$.

If $(\rho_i - \rho_{i-1})(\kappa_i - \kappa_{i-1}) > 0$, the rays $\mathfrak{Y}_{i-1}$ and $\mathfrak{Y}_i$ cut the rectangle $\mathscr{R}_i$ into three pieces, and it may happen that $\widehat{M}_i$ lies outside the cone, which violates the desired min-max principle. □

To know what should be done for the case $(\rho_i - \rho_{i-1})(\kappa_i - \kappa_{i-1}) > 0$, we introduce

$$Y_i^m = \min\{Y_{i-1}, Y_i\}, \qquad Y_i^M = \max\{Y_{i-1}, Y_i\}, \tag{3.6}$$

and seek sufficient conditions at a given cell i under the assumption $\lambda < 1$.

LEMMA 3.3. *For a given cell i, if*

$$Y_i^M D_i^\rho + \tfrac{2}{\lambda}(\kappa_i - Y_i^M \rho_i) \le D_i^\kappa \le Y_i^m D_i^\rho + \tfrac{2}{\lambda}(\kappa_i - Y_i^m \rho_i), \tag{3.7a}$$

$$Y_i^m D_{i-1}^\rho - \tfrac{2}{1-\lambda}(\kappa_{i-1} - Y_i^m \rho_{i-1}) \le D_{i-1}^\kappa \le Y_i^M D_{i-1}^\rho - \tfrac{2}{1-\lambda}(\kappa_{i-1} - Y_i^M \rho_{i-1}), \tag{3.7b}$$

then $Y_i^m \le \widehat{Y}_i \le Y_i^M$.

Proof. A straightforward calculation shows that

$$\widehat{\kappa}_i - Y_i^m \widehat{\rho}_i = (1-\lambda)A_i^m + \lambda B_{i-1}^m, \qquad \widehat{\kappa}_i - Y_i^M \widehat{\rho}_i = (1-\lambda)A_i^M + \lambda B_{i-1}^M, \tag{3.8}$$

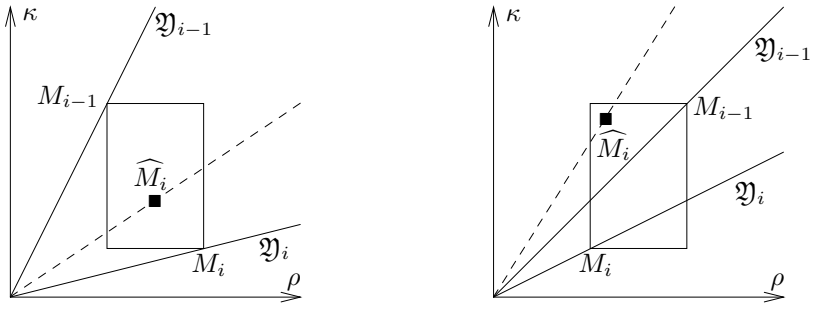

FIG. 3.1. *Geometric analysis of the min-max principle for the fraction problem.*

with

$$A_i^m = (\kappa_i - Y_i^m \rho_i) - \tfrac{\lambda}{2}(D_i^\kappa - Y_i^m D_i^\rho) \tag{3.9a}$$
$$A_i^M = (\kappa_i - Y_i^M \rho_i) - \tfrac{\lambda}{2}(D_i^\kappa - Y_i^M D_i^\rho) \tag{3.9b}$$
$$B_{i-1}^m = (\kappa_{i-1} - Y_i^m \rho_{i-1}) + \tfrac{1-\lambda}{2}(D_{i-1}^\kappa - Y_i^m D_{i-1}^\rho) \tag{3.9c}$$
$$B_{i-1}^M = (\kappa_{i-1} - Y_i^M \rho_{i-1}) + \tfrac{1-\lambda}{2}(D_{i-1}^\kappa - Y_i^M D_{i-1}^\rho). \tag{3.9d}$$

In order to ensure $\widehat{\kappa}_i - Y_i^m \widehat{\rho}_i \ge 0$ and $\widehat{\kappa}_i - Y_i^M \widehat{\rho}_i \le 0$, our strategy consists in splitting the summands involved in (3.8). By forcibly imposing

$$A_i^m \ge 0, \quad A_i^M \le 0, \quad B_{i-1}^m \ge 0, \quad B_{i-1}^M \le 0, \tag{3.10}$$

we end up with the set of inequalities (3.7). □

THEOREM 3.4. *Given an initial choice $(\widetilde{D}_i^\rho, \widetilde{D}_i^\kappa)$ in accordance with (3.3), let $\mathscr{Y}_i \subset \mathbb{R}^2$ be the set of all pairs (D_i^ρ, D_i^ρ) subject to the 8 linear inequality constraints*

$$[\widetilde{D}_i^\rho]^- \le D_i^\rho \le [\widetilde{D}_i^\rho]^+, \quad [\widetilde{D}_i^\kappa]^- \le D_i^\kappa \le [\widetilde{D}_i^\kappa]^+, \quad \mathfrak{m}_i(D_i^\rho) \le D_i^\kappa \le \mathfrak{M}_i(D_i^\rho), \tag{3.11}$$

with

$$\mathfrak{m}_i(D_i^\rho) = \max\{ Y_i^M D_i^\rho + \tfrac{2}{\lambda}(\kappa_i - Y_i^M \rho_i) \,;\, Y_{i+1}^m D_i^\rho - \tfrac{2}{1-\lambda}(\kappa_i - Y_{i+1}^m \rho_i) \} \tag{3.12a}$$
$$\mathfrak{M}_i(D_i^\rho) = \min\{ Y_i^m D_i^\rho + \tfrac{2}{\lambda}(\kappa_i - Y_i^m \rho_i) \,;\, Y_{i+1}^M D_i^\rho - \tfrac{2}{1-\lambda}(\kappa_i - Y_{i+1}^M \rho_i) \}. \tag{3.12b}$$

For all $i \in \mathbb{Z}$, define

$$(D_i^\rho, D_i^\kappa) = \begin{cases} (\widetilde{D}_i^\rho, \widetilde{D}_i^\kappa) & \text{if } \widetilde{D}_i^\rho \widetilde{D}_i^\kappa < 0 \\ \Pi_{\mathscr{Y}_i}(\widetilde{D}_i^\rho, \widetilde{D}_i^\rho) & \text{otherwise} \end{cases} \tag{3.13}$$

where $\Pi_{\mathscr{Y}_i}(.)$ denotes the projection onto the convex set $\mathscr{Y}_i \subset \mathbb{R}^2$. Then, we have the min-max principles

$$\widehat{\rho}_i \in \lfloor \rho_{i-1}, \rho_i \rceil, \qquad \widehat{\kappa}_i \in \lfloor \kappa_{i-1}, \kappa_i \rceil, \qquad \widehat{Y}_i \in \lfloor Y_{i-1}, Y_i \rceil, \tag{3.14}$$

at every cell i when updating (ρ, κ) with scheme (3.2).

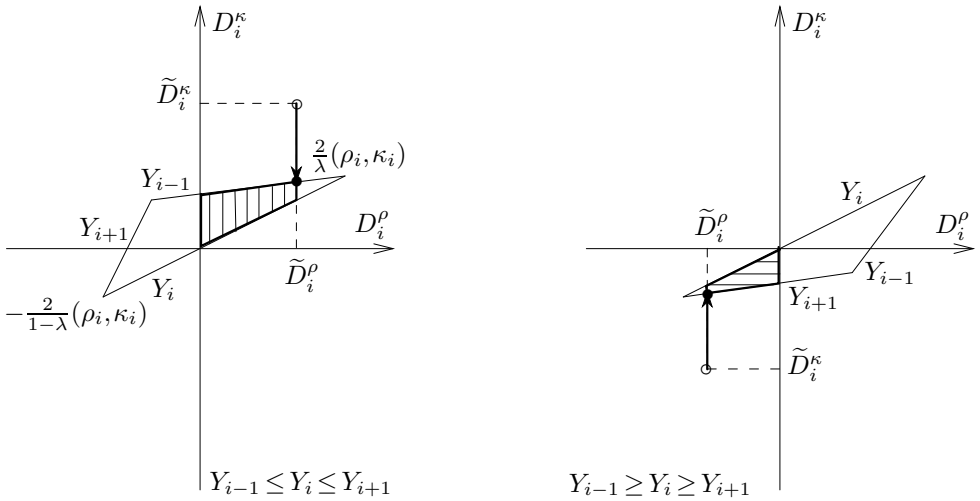

FIG. 3.2. *Projection of $(\widetilde{D}_i^\rho,\widetilde{D}_i^\kappa)$ onto the convex set $\mathscr{Y}_i$ for the fraction problem.*

Proof. The proof is similar to that of Theorem 2.4. □

The practical implementation of the projection onto $\mathscr{Y}_i$ in this problem can be done via explicit formulae. Figure 3.2 displays a few situations for a locally increasing or decreasing behavior of Y. It can be readily proven that the constraints $\mathfrak{m}_i(D_i^\rho)\le D_i^\kappa\le\mathfrak{M}_i(D_i^\rho)$ correspond in reality to a triangle in the (D_i^ρ,D_i^κ)-plane. The slopes of its sides are Y_{i-1}, Y_i and Y_{i+1}. The side with slope Y_i passes through the origin and connects the points $-\frac{2}{1-\lambda}(\rho_i,\kappa_i)$ and $\frac{2}{1-\lambda}(\rho_i,\kappa_i)$.

Should a local extremum occurs, i.e., $(Y_{i-1}-Y_i)(Y_{i+1}-Y_i)>0$, this triangle degenerates into the segment joining these two points. Hence, $D_i^\kappa=Y_iD_i^\rho$ whenever the projection operator $\Pi_{\mathscr{Y}_i}$ is activated, and we formally have $D_i^Y=(D_i^\kappa-Y_iD_i^\rho)/\rho_i=0$. This testifies to a clipping mechanism on Y.

3.2. Variable velocity field The setting is identical to that of the sum problem. Introduce the local bounds

$$\begin{aligned}Y_i^m=&\mathbb{1}_{\{S(i)=\mathrm{I}\}}\ \min\{Y_{i-1},Y_i\} &(3.15a)\\ &+\mathbb{1}_{\{S(i)=\mathrm{II}\}}\min\{Y_{i+1},Y_i\}+\mathbb{1}_{\{S(i)=\mathrm{III}\cup\mathrm{IV}\}}\min\{Y_{i-1},Y_i,Y_{i+1}\}\\ Y_i^M=&\mathbb{1}_{\{S(i)=\mathrm{I}\}}\ \max\{Y_{i-1},Y_i\} &(3.15b)\\ &+\mathbb{1}_{\{S(i)=\mathrm{II}\}}\max\{Y_{i+1},Y_i\}+\mathbb{1}_{\{S(i)=\mathrm{III}\cup\mathrm{IV}\}}\max\{Y_{i-1},Y_i,Y_{i+1}\}.\end{aligned}$$

Once all the Y^m and Y^M have been computed over the domain, we consider the set $\mathscr{Y}_i$ of all pairs (D_i^ρ,D_i^κ) that satisfy:

- For case I (left-to-right propagation)

$$[\widetilde{D}_i^\rho]^-\le D_i^\rho\le[\widetilde{D}_i^\rho]^+,\quad[\widetilde{D}_i^\kappa]^-\le D_i^\kappa\le[\widetilde{D}_i^\kappa]^+,\quad\mathfrak{m}_i(D_i^\rho)\le D_i^\kappa\le\mathfrak{M}_i(D_i^\rho),\quad(3.16)$$

with

$$\begin{aligned}\mathfrak{m}_i(D_i^\rho)=&\max\{\,Y_i^MD_i^\rho+\Phi_i(\kappa_i-Y_i^M\rho_i)\,;\,Y_{i+1}^mD_i^\rho-\Phi_i(\kappa_i-Y_{i+1}^m\rho_i)\,\}&(3.17a)\\ \mathfrak{M}_i(D_i^\rho)=&\min\{\,Y_i^mD_i^\rho+\Phi_i(\kappa_i-Y_i^m\rho_i)\,;\,Y_{i+1}^MD_i^\rho-\Phi_i(\kappa_i-Y_{i+1}^M\rho_i)\,\}.&(3.17b)\end{aligned}$$

- For case II (right-to-left propagation)

$$[\widetilde{D}_i^\rho]^- \le D_i^\rho \le [\widetilde{D}_i^\rho]^+, \quad [\widetilde{D}_i^\kappa]^- \le D_i^\kappa \le [\widetilde{D}_i^\kappa]^+, \quad \mathfrak{m}_i(D_i^\rho) \le D_i^\kappa \le \mathfrak{M}_i(D_i^\rho), \tag{3.18}$$

with

$$\mathfrak{m}_i(D_i^\rho) = \max\{\, Y_i^M D_i^\rho + \Phi_i(\kappa_i - Y_i^M \rho_i)\,; Y_{i-1}^m D_i^\rho - \Phi_i(\kappa_i - Y_{i-1}^m \rho_i)\,\} \tag{3.19a}$$

$$\mathfrak{M}_i(D_i^\rho) = \min\{\, Y_i^m D_i^\rho + \Phi_i(\kappa_i - Y_i^m \rho_i)\,; Y_{i-1}^M D_i^\rho - \Phi_i(\kappa_i - Y_{i-1}^M \rho_i)\,\}. \tag{3.19b}$$

- For case III (source)

$$[\widetilde{D}_i^\rho]^- \le D_i^\rho \le [\widetilde{D}_i^\rho]^+, \quad [\widetilde{D}_i^\kappa]^- \le D_i^\kappa \le [\widetilde{D}_i^\kappa]^+, \quad \mathfrak{m}_i(D_i^\rho) \le D_i^\kappa \le \mathfrak{M}_i(D_i^\rho), \tag{3.20}$$

with

$$\begin{aligned} \mathfrak{m}_i(D_i^\rho) = \max\{\, & Y_{i-1}^M D_i^\rho + \Phi_i(\kappa_i - Y_{i-1}^M \rho_i); Y_i^m D_i^\rho - \Phi_i(\kappa_i - Y_i^m \rho_i); \\ & Y_i^M D_i^\rho + \Phi_i(\kappa_i - Y_i^M \rho_i)\,; Y_{i+1}^m D_i^\rho - \Phi_i(\kappa_i - Y_{i+1}^m \rho_i)\,\} \end{aligned} \tag{3.21a}$$

$$\begin{aligned} \mathfrak{M}_i(D_i^\rho) = \min\{\, & Y_{i-1}^m D_i^\rho + \Phi_i(\kappa_i - Y_{i-1}^m \rho_i); Y_i^m D_i^\rho + \Phi_i(\kappa_i - Y_i^m \rho_i); \\ & Y_i^M D_i^\rho - \Phi_i(\kappa_i - Y_i^M \rho_i)\,; Y_{i+1}^M D_i^\rho - \Phi_i(\kappa_i - Y_{i+1}^M \rho_i)\,\}. \end{aligned} \tag{3.21b}$$

- For case IV (sink), $\mathscr{Y}_i = \mathbb{R}^2$.

THEOREM 3.5. *Given an initial choice $(\widetilde{D}_i^\rho, \widetilde{D}_i^\kappa)$ in accordance with (3.3), (2.19), let $\mathscr{Y}_i \subset \mathbb{R}^2$ be the convex set introduced above. For all $i \in \mathbb{Z}$, define*

$$(D_i^\rho, D_i^\kappa) = \begin{cases} (\widetilde{D}_i^\rho, \widetilde{D}_i^\kappa) & \text{if } \widetilde{D}_i^\rho \widetilde{D}_i^\kappa < 0 \\ \Pi_{\mathscr{Y}_i}(\widetilde{D}_i^\rho, \widetilde{D}_i^\kappa) & \text{otherwise} \end{cases} \tag{3.22}$$

where $\Pi_{\mathscr{Y}_i}(.)$ denotes the projection onto the convex set $\mathscr{Y}_i \subset \mathbb{R}^2$. Then, under assumptions (2.18) and (2.20), we have the min-max principle (2.17) for $\psi = \rho, \kappa$ and Y at every cell i when updating (ρ, κ) with scheme (2.15).

Proof. The proof is similar to that of Theorem 2.5. □

Again, we recommend a minimization subroutine to perform the projection.

REMARK 3.6. *In the (D_i^ρ, D_i^κ)-plane, let* A *and* B *be the points located at*

$$\mathtt{A} = -\frac{1}{\Phi_i}(\rho_i, \kappa_i), \qquad \mathtt{B} = \frac{1}{\Phi_i}(\rho_i, \kappa_i).$$

It can be shown that, the set of points defined by the inequalities $\mathfrak{m}_i(D_i^\rho) \le D_i^\kappa \le \mathfrak{M}_i(D_i^\rho)$ in cases I, II *and* III *is the segment* AB *if $(Y_i - Y_{i-1})(Y_i - Y_{i+1}) \ge 0$. For $(Y_i - Y_{i-1})(Y_i - Y_{i+1}) < 0$, this domain always contain the triangle* ABC*, in which the slope of* (AC) *is Y_{i+1} and the slope of* (CB) *is Y_{i-1}. Therefore, if we accept to project onto a smaller convex set, it is possible to find the new slopes by explicit formulae. The price to be paid for is a larger amount of dissipation.*

3.3. Numerical results In Figures 3.3 and 3.4, we compare the results of 3 different schemes and the exact solution for an experiment over a positive velocity field. The initial data have been tailored so that ρ and and β are both increasing, therefore we have $(\rho_i - \rho_{i-1})(\kappa_i - \kappa_{i-1}) \ge 0$ for all cell i in the domain. The curves for the first-order scheme are very much smeared out. Those for the two second-order schemes are in very good agreement with the exact solution. The labels "old second-order" and "new second-order" have the same meaning as in §2.3. We see that in the vicinity of $x = 1100$, the old second-order scheme does not comply with the min-max principle on the control variable Y. As for the new second-order scheme, it does not exhibit any oscillation on Y, as testified by the lower panel of Figure 3.4.

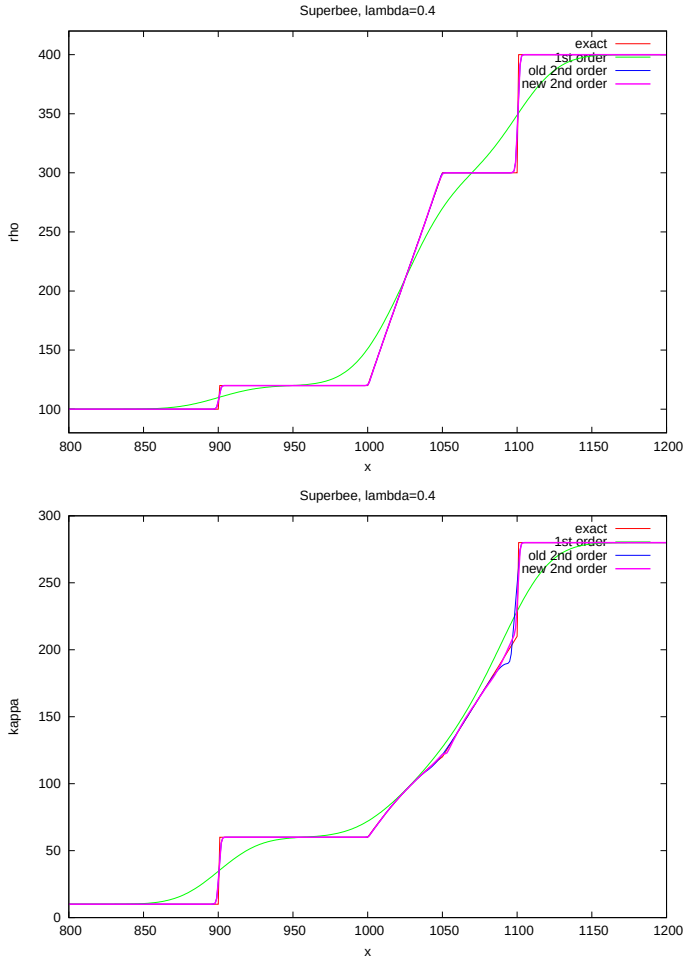

FIG. 3.3. *Main variables ρ (upper panel) and κ (lower panel) for the fraction problem.*

4. The general problem We go back to the problem stated in the Introduction. The ideas presented for the sum problem and the fraction problem can be carried over to the case of several control variables $\mathbf{G}$, each of them being a first-order rational fraction with respect to $\mathbf{\Psi} \in \mathbb{R}_+^P$, that is,

$$G^q(\mathbf{\Psi}) = \frac{a_0^q + a_1^q \psi^1 + \ldots + a_P^q \psi^P}{b_0^q + b_1^q \psi^1 + \ldots + b_P^q \psi^P}, \tag{4.1}$$

with $b_p^q \geq 0$ for all $1 \leq p \leq P$, $1 \leq q \leq Q$. This class of homographic functions is wide enough to represent a vast majority of control variables in real-life applications.

4.1. Uniform velocity Assuming $u > 0$, we define

$$(\psi_i^q)^m = \min\{\psi_{i-1}^q, \psi_i^q\}, \qquad (\psi_i^q)^M = \max\{\psi_{i-1}^q, \psi_i^q\}, \tag{4.2}$$

and

$$(G_i^q)^m = \min\{G_{i-1}^q, G_i^q\}, \qquad (G_i^q)^M = \max\{G_{i-1}^q, G_i^q\}. \tag{4.3}$$

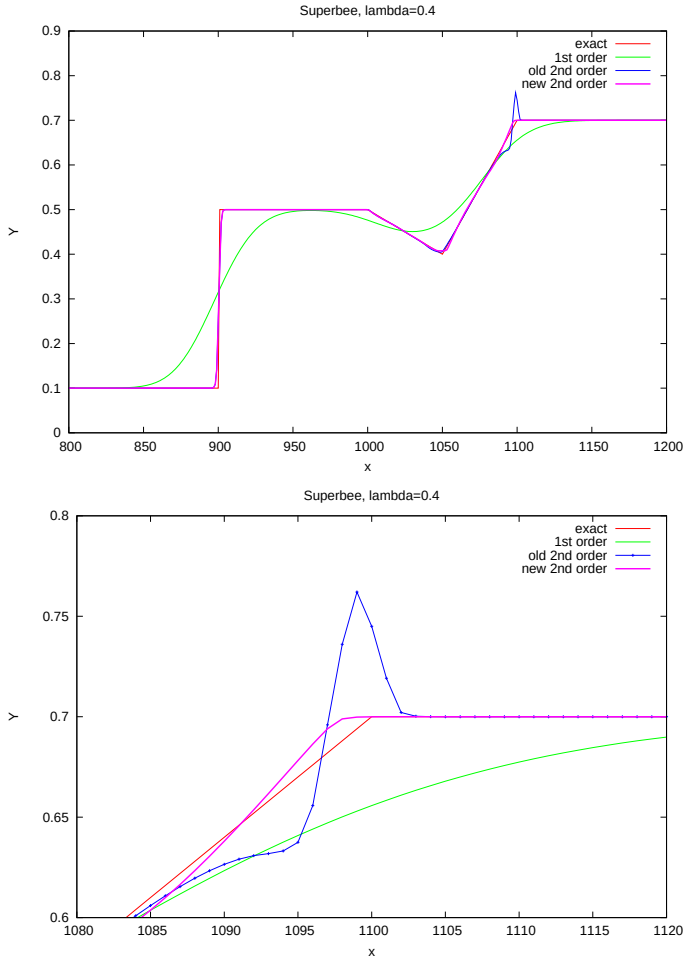

FIG. 3.4. *Control variable $Y=\kappa/\rho$ for the fraction problem.*

Our objective is to find the slopes $\mathbf{D}_i=(D_i^1,\dots,D_i^P)$, which should be as close as possible to the initial slopes $\widetilde{\mathbf{D}}_i=(\widetilde{D}_i^1,\dots,\widetilde{D}_i^P)$ computed component-wise by a standard limiter function, so that by updating $\mathbf{\Psi}$ with (1.5), we have not only

$$(\psi_i^q)^m \le \widehat{\psi}_i^q \le (\psi_i^q)^M, \tag{4.4}$$

but also

$$(G_i^q)^m \le \widehat{G}_i^q = G^q(\widehat{\mathbf{\Psi}}_i) \le (G_i^q)^M. \tag{4.5}$$

Getting rid of the denominator in G^q, the above condition can be cast into two linear inequalities involving $(\mathbf{\Psi}_{i-1},\mathbf{\Psi}_i)$ and $(\mathbf{D}_{i-1},\mathbf{D}_i)$. The splitting strategy enables us to break these inequalities into local conditions which do not couple $\mathbf{D}_{i-1}$ and $\mathbf{D}_i$. These conditions, once gathered, express that we must project the initial guess $\widetilde{\mathbf{D}}_i$ onto a convex set $\mathscr{G}_i\subset\mathbb{R}^P$ defined by $2P$ bound constraints (to ensure monotonicity on $\mathbf{\Psi}$) and $4Q$ non-trivial linear inequalities (to ensure monotonicity on $\mathbf{G}$).

To carry out this projection, we reformulate the projection operator as a quadratic minimization problem

$$\min_{\mathbf{D}_i \in \mathscr{G}_i} \tfrac{1}{2} \|\mathbf{D}_i - \widetilde{\mathbf{D}}_i\|^2 \tag{4.6}$$

subject to linear inequality constraints. We recall that by virtue of Hilbert's theorem about projection onto a convex non-empty set, there is a unique solution to problem (4.6). In the context of the applications we have in mind, the Euclidean norm does make sense, insofar as the components of Ψ are homogeneous to a density. We advocate the use of an existing subroutine, e.g., the QL algorithm by Schittkowski [6], the advantage of which lies in its fast convergence. Moreover, it can be initialized with $\mathbf{D}_i = \widetilde{\mathbf{D}}_i$, which is not necessarily a feasible point.

Before launching the optimization procedure, however, we have to carefully determine the regions in the $\mathbf{D}_i$-space for which the min-max principle on $\mathbf{G}$ is automatically guaranteed and for which there is no need to perform projection (for the sum problem, this region is $D_i^\alpha D_i^\beta > 0$ and for the fraction problem, this is $D_i^\rho D_i^\kappa < 0$). This crucial preliminary step is meant to maintain sharp profiles. It can only be done on a case-by-case basis.

4.2. Variable velocity field The ideas remain the same as in the uniform velocity case, but the calculations are trickier. On one hand, the definitions of the local bounds depend on the sign configuration at the edges of each cell. On the other hand, the inequalities to be split now involve $\mathbf{D}_{i-1}, \mathbf{D}_i$ and $\mathbf{D}_{i+1}$. As a consequence, after imposing positivity or negativity to the summands separately, we end up with more than $4Q$ non-trivial combinations for case III (source). Nevertheless, this is not a difficulty because the hard part of the job is done by the optimization subroutine.

The extra time incurred by the latter depends on the size of P and Q. Numerical experiments reveal that for a typical multi-specie flow model ($P \approx 10$, $Q \approx 5$), such as in [1,8], the CPU ratio never exceeds 2. Besides, we have to be aware of the fact that the remap phase contributes little to the overall computational time of the nonlinear Euler code. From this global point of view, the reward brought by the min-max principle on the main and control variables is worth an increase by a factor 2 in the CPU time of the remap phase, which is not really significant!

4.3. Selected examples In addition to the sum problem and the fraction problem, we have successfully applied the new slope-limiting method to the following examples, in which there are two control variables. Since the conclusion is the same as before, we simply state the problem and do not show the curves.

4.4. The fraction-difference problem Consider $\mathbf{\Psi} = (\rho, \kappa) \in \mathbb{R}_+^2$ and

$$\mathbf{G}(\mathbf{\Psi}) = (Y, \eta) = \left(\frac{\kappa}{\rho}, \rho - \kappa \right). \tag{4.7}$$

As explained in §3, we are in a two-phase flow model, the total density of which is ρ and the gas density of which is κ. The ratio $Y = \kappa/\rho$ represents the gas mass fraction, while the difference $\eta = \rho - \kappa$ is equal to the liquid density.

4.5. The two-sum four-specie problem Consider $\mathbf{\Psi} = (\alpha, \beta, \gamma, \delta) \in \mathbb{R}_+^4$ and

$$\mathbf{G}(\mathbf{\Psi}) = (e, f) = (7\alpha + \gamma, 2\beta + 2\gamma + \delta) \tag{4.8}$$

The densities α, β, γ, δ respectively correspond to the four species C_7H_{16}, O_2, CO_2, H_2O, related to each other through the reversible chemical reaction

$$C_7H_{16} + 11O_2 \rightleftarrows 7CO_2 + 8H_2O \tag{4.9}$$

which takes place at the known rate

$$R(\alpha,\beta,\gamma,\delta) = K_+\alpha\beta^{11} - K_-\gamma^7\delta^8. \tag{4.10}$$

This implies the evolution equations

$$D_t\alpha = - \quad R(\alpha,\beta,\gamma,\delta) \tag{4.11a}$$
$$D_t\beta = -11R(\alpha,\beta,\gamma,\delta) \tag{4.11b}$$
$$D_t\gamma = \quad 7R(\alpha,\beta,\gamma,\delta) \tag{4.11c}$$
$$D_t\delta = \quad 8R(\alpha,\beta,\gamma,\delta), \tag{4.11d}$$

where D_t denotes the total derivative $\partial_t + u\partial_x$. From (4.11), it can be inferred that

$$D_t e = D_t f = 0, \tag{4.12}$$

which highlights e and f as control variables. In the present case, e is the carbon tracer, and f the oxygen tracer.

5. Conclusion We hope the slope-reconstruction methodology proposed in this paper, based on a rigorous analysis while being not too much expensive, will be helpful to the practitioners who have to daily face similar problems. Current works are in progress in order to extend this approach to multi-dimensional problems.

Acknowledgement. The author is grateful to Frédéric Coquel and Bruno Scheurer for helpful discussions and for their valuable comments on the manuscript.

REFERENCES

[1] O. Colin and A. Benkenida, *The 3-zones extended coherent flame model (ECFM3Z) for computing premixed/diffusion combustion*, Oil & Gas Sci. Tech., 59 (2004), pp. 593–609.
[2] F. Coquel, Q. L. Nguyen, M. Postel, and Q. H. Tran, *Entropy-satisfying relaxation method with large time-steps for Euler IBVPs*, Submitted, (2007).
[3] C. W. Hirt, A. A. Amsden, and J. L. Cook, *An arbitrary Lagrangian-Eulerian computing method for all flow speeds*, J. Comput. Phys., 14 (1974), pp. 227–253.
[4] R. J. LeVeque, *Numerical Methods for Conservation Laws*, Lectures in Mathematics, ETH Zürich, Birkhäuser Verlag, Berlin, 1992.
[5] ———, *Finite Volume Methods for Hyperbolic Problems*, vol. 31 of Cambridge Texts in Applied Mathematics, Cambridge University Press, Cambridge, 2002.
[6] K. Schittkowski, *NLPQL: A Fortran subroutine solving constrained nonlinear programming problems*, Ann. Oper. Res., 5 (1986), pp. 485–500.
[7] P. K. Sweby, *High resolution schemes using flux limiters for hyperbolic conservation laws*, SIAM J. Numer. Anal., 21 (1984), pp. 995–1011.
[8] Q. H. Tran and B. Scheurer, *High-order monotonicity-preserving compact schemes for linear advection on 2-d irregular meshes*, J. Comput. Phys., 175 (2002), pp. 454–486.
[9] B. van Leer, *Towards the ultimate conservative difference scheme V: A second-order sequel to Godunov's method*, J. Comput. Phys., 32 (1979), pp. 101–136.
[10] W. B. VanderHeyden and B. A. Kashiwa, *Compatible fluxes for van Leer advection*, J. Comput. Phys., 146 (1998), pp. 1–28.